A Scientific Knowledge Discovery and Data Mining Process Model for Metabolomics

BaniMustafa, Ahmed; Hardy, Nigel

Published in:
IEEE Access

DOI:
10.1109/ACCESS.2020.3039064

Publication date:
2020

Citation for published version (APA):
BaniMustafa, A., & Hardy, N. (2020). A Scientific Knowledge Discovery and Data Mining Process Model for Metabolomics. *IEEE Access*, *8*, 209964 - 210005. https://doi.org/10.1109/ACCESS.2020.3039064

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 22, 2020, accepted October 15, 2020, date of publication November 18, 2020,
date of current version December 3, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3039064

A Scientific Knowledge Discovery and Data Mining Process Model for Metabolomics

AHMED BANIMUSTAFA[1], (Member, IEEE), AND NIGEL HARDY[2]
[1]Department of Software Engineering, ISRA University, Amman 11622, Jordan
[2]Department of Computer Science, Aberystwyth University, Aberystwyth SY23 3DB, U.K.
Corresponding author: Ahmed BaniMustafa (banimustafa@acm.org)

ABSTRACT This work presents a scientific data mining process model for metabolomics that provides a systematic and formalised framework for guiding and performing metabolomics data analysis in a justifiable and traceable manner. The process model is designed to promote the achievement of the analytical objectives of metabolomics investigations and to ensure the validity, interpretability and reproducibility of their results. It satisfies the requirements of metabolomics data mining, focuses on the contextual meaning of metabolomics knowledge, and addresses the shortcomings of existing data mining process models, while paying attention to the practical aspects of metabolomics investigations and other desirable features. The process model development involved investigating the ontologies and standards of science, data mining and metabolomics and its design was based on the principles, best practices and inspirations from Process Engineering, Software Engineering, Scientific Methodology and Machine Learning. A software environment was built to realise and automate the process model execution and was then applied to a number of metabolomics datasets to demonstrate and evaluate its applicability to different metabolomics investigations, approaches and data acquisition instruments on one hand, and to different data mining approaches, goals, tasks and techniques on the other. The process model was successful in satisfying the requirements of metabolomics data mining and can be generalised to perform data mining in other scientific disciplines.

INDEX TERMS Data mining, bioinformatics, computational biology, knowledge discovery, machine learning, metabolomics data analysis, process engineering, software engineering.

I. INTRODUCTION

Metabolomics is a newly emerging field that has a huge potential in a wide range of domains and applications such as genetics, medicine, nutrition, agriculture, and environment. Yet, metabolomics data is complex and heterogeneous and it tells nothing without proper analysis and interpretation. In addition, the analysis of metabolomics data is often approached in an informal manner and without explicit theoretical justification. It is usually performed based on analysts' hunches and often influenced by his/her personal preference, background and experience. There is a general agreement in the metabolomics community that metabolomics does not only depend on the advance of chemical analysis techniques and data acquisition instruments, but also on advances in computational and data analysis methods.

The scientific nature of metabolomics investigations requires attention to issues, including traceability of the analysis procedures, the justification for decisions and the reproducibility of results. The analysis process must be governed by clear guidance and should be carried out through a systematic approach and pre-described steps. Therefore, the analysis of metabolomics data must be consistent with the cycle of knowledge which supports both deductive and inductive knowledge acquisition, while the first focuses on inducing knowledge from observations, the latter focuses on deducing knowledge by testing an hypothesis [1]. In addition, the analysis must also be driven by well-defined objectives and must consider the nature and quality of the acquired metabolomics dataset. All activities in the analysis process must be planned and validated and the practical aspects of metabolomics data analysis must also be considered. The results of the analysis must be presented in an interpretable and evaluable fashion and the final outcomes must be deployable in order to be utilised either for further analysis or to be compared with

The associate editor coordinating the review of this manuscript and approving it for publication was Dongxiao Yu.

the results that are generated from future or similar investigations.

To fulfill these requirements and address the above issues, this paper presents a scientific knowledge discovery and data mining process model for metabolomics that provides a systematic framework for conducting metabolomics data analysis and guides its execution process. The development of the process started with investigating the fundamentals of data mining and metabolomics investigations and involved identifying metabolomics requirements that covered metabolomics data acquisition, preprocessing, pretreatments, data analysis and results interpretation. The process design was then laid out based on the ontologies and standards of scientific investigations and data mining and was then enhanced and improved based on inspirations from a number of interdisciplinary fields that considered the fundamentals, principles and best practices of process engineering, software engineering, machine Learning and scientific methodology, while paying attention to the practical aspects of metabolomics investigations, beside a number of other desirable features that was reported by data mining and metabolomics practitioners in the literature. In addition, a critical review of existing data mining process models was conducted which concluded that they suffer from flaws and shortcomings which makes them inappropriate for metabolomics data mining due to their generic nature and to their lack of scientific orientation and attention to important issues concern metabolomics data mining such as justifiability, traceability and reproducibility [2]. The analysis also found that most of these process models suffer from issues regarding their structure, layout design and flow based on process engineering principles and best practices including cohesion, abstraction, modularization, notation, content, instantiation and understandability. They lack support for practical aspects such as: management, quality, standardisation and human interaction and also do not pay attention to desirable features of reported in both metabolomics and data mining such as: visualisation, data exploration, knowledge presentation and automation. Reviews and surveys that were reported in the literature highlight some of these issues and confirms the results of the analysis [3]–[8], while others made efforts to address these issues either through extending existing process that are applicable in specific domains such as health domain [9] and information management [10], or through proposing generic [11] and specific process models that are dedicated to specific domains such as: health care [12] and data analytics [13].

MeKDDaM is designed to be generic, so it covers all types of metabolomics investigations and studies, regardless of their approaches or data acquisition instruments. It is, however, customizable, so that it satisfies the special needs of a particular metabolomics data analysis. It supports both hypothesis-driven and data-driven knowledge discovery and it covers all data mining approaches, goals, tasks and techniques. The process model focuses on the contextual aspects of knowledge discovery and on satisfying the fundamentals principles of the scientific methodology. It allows execution of data mining procedures in a systematic and justifiable manner and support generation of traceable results which contributes towards the reproducibility of metabolomics data analysis. The process model was realised and automated using software that was designed to guide the execution of the process (MeKDDaM-SAGA). The software was developed based on object-oriented Software Engineering methodology and constructed in Java [14]. The development of the software is described in an ArXiv preprint [15]. MeKDDaM-SAGA is freely available as an open-source software at GitHub software repository (https://github.com/banimustafa/MekDDaM-SAGA) [16]. MeKDDaM was then applied to four metabolomics applications to demonstrate its applicability and to evaluate its execution in the context of its development requirements [14]. The demonstration applications are reported in a preprint article that is available on an ArXiv [14].

The following section provides an overview of the fundamental concepts of both metabolomics and data mining, while Section III provides a description for the process model development methodology. Section IV provides a description of the proposed process model, while Section IV-H provides an evaluation of the proposed model in the light of the process execution and demonstrated applications.

II. BACKGROUND

In this section we first discuss the fundamental concepts and experimental design of metabolomics investigations that are related to the data acquisition techniques, data preprocessing and data pretreatment procedures which are related to the nature and quality of metabolomics datasets and their associated meta-data. Second, we discuss the fundamentals of knowledge discovery and data mining and cover its concepts, approaches, goals and tasks in addition to applied processes and techniques.

A. METABOLOMICS DATA ANALYSIS

Metabolomics is defined as *the study of all low molecule weight chemicals (metabolites) which are involved in metabolism, either as an end product or as necessary chemicals for metabolism* [17]–[19] while **metabolites** are the low-weight chemical compounds with molecular-weight less than 1000Da. On the other hand, metabolic approaches refer to the way in which metabolomics chemical analysis is performed and to the nature of its results. Below is a list of metabolic approaches:

- **Metabolite Profiling**: Identifying and quantifying a predefined set of metabolites which belong to particular subgroups of chemical classes or involved in a particular biochemical pathway [19]–[21]. The measured metabolites are stored in a profile as pairs of metabolites signals and their associated intensity or concentration.
- **Metabolite Target Analysis**: A form of metabolite profiling that provides a qualitative and quantitative

analysis of a particular set of metabolites in a specific metabolic reaction [19], biological system or biochemical pathway such as enzymes [20], [22]. It measures the concentration of the signals of the targeted metabolites and quantifies their levels [23].

- **True Metabolomics**: provides unbiased and comprehensive measurements of the overall metabolome under a particular condition [20], [22], [24]. It measures the concentrations of the metabolites by identifying and quantifying all of the existing metabolites [23].
- **Metabolic Fingerprinting**: A rapid, global, high throughput analysis which aims at discovering patterns and classifying samples [19], [24] without the need to identify or quantify the metabolites involved [19], [23].

Metabolomics analysis is a complex process, which requires careful experimental design. The data captured by the data acquisition techniques is complex and heterogeneous. It may include hundreds of variables that need to be carefully analysed in order to complement other ''omics'' techniques or to answer a particular metabolomics investigation question. Metabolomics data analysis is influenced by the goals of the metabolomics investigation, the applied metabolic approach, and the design of the metabolomics study and its subsequent assays. These influence the nature and quality of the data and its required preprocessing and pretreatment procedures. The design of metabolomics studies is driven by the analytical objectives of metabolomics investigation which drive the selection of the sampling method, metabolic approach, data acquisition techniques and instruments [25], [26]. Figure 1 illustrates the design of metabolomics studies and the types of data it generates.

Metabolomics data acquisition employs a number of analysis instruments that are used either alone or in combination with other techniques in order to analyze bio-samples and generate metabolomics datasets. These techniques belong to four main groups: (1) Liquid and Gas Chromatography Separation [19], [27]; (2) Mass Spectrometry [28]; (3) Optical [25]; and (4) Nuclear Magnetic Resonance techniques [29]. The selection of these data acquisition techniques is usually driven by the aims of the study, the research question, the goal of the investigation and the metabolomics approach applied. This choice influences the nature and quality of the data and its preprocessing and pretreatment procedures. It also influences the data analysis objectives which can be achieved. The nature of the data generated during data acquisition varies depending on the analytical instrument or the combination of analytical instruments used and the data transformations applied to the data [30]. In addition, meta-data which covers information regarding the instruments, their settings, runs and adjustments must also be collected. Data exporting facilities are used in order to convert the data into independent formats that enable its accessing by external software. Data models are generally used in order to improve the quality of the collected data and enable storing and accessing data [25]. ArMet [31], FuGe [32], RSBI(ISA-TAB) [33] are examples of these models which also consider the storage of meta-data.

B. KNOWLEDGE DISCOVERY AND DATA MINING

Metabolomics knowledge discovery has an important potential for understanding metabolic phenotyping, metabolic fluxes, intermediate metabolic pathways and metabolites networking in addition to complementing the work of other ''omics'' and completing the overall picture of systems biology [22]. Thus, metabolomics can harness advances in machine learning and data mining in order to overcome the complexity of metabolomics analysis and to guide the execution of the analysis process. Knowledge discovery has achieved noticeable success in a variety of applications, widely in business and on a narrower scale in science, making use of advances in machine learning and other computational methods. Data mining can be used to tackle the complexity of metabolomics data and uncover its knowledge. Recently, a growing number of data mining applications have been reported in metabolomics [34]–[38].

Knowledge Discovery is defined as the ''*non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data*'' [39], while data mining is considered a step in knowledge discovery. Knowledge discovery has two aspects: data mining techniques and process models. The first concerns building a valid model using a wide variety of techniques adopted from several fields including statistics, machine learning, pattern discovery and other computational methods, while the latter concerns the set of activities required for applying the technique and ultimately extracting knowledge.

Data mining can be performed using either a data-driven or a hypothesis-driven approach. The data-driven approach aims at uncovering novel or interesting knowledge in the form of patterns, trends, association or other kinds of relationships in the data regardless of the original purpose of its acquisition [40]. It is suitable for exploratory objectives, when few or no assumptions are available regarding what is expected in the data. On the other hand, hypothesis-driven approach aims to test a preexisting hypothesis or assertion and is usually carried out for confirmatory or exploratory reasons [41].

Data mining goals are related to the reason for applying data mining techniques, while data mining tasks are related to the purpose that they seek to achieve. Data mining techniques perform a specific task that is linked to a goal. Discovery-oriented goals are aimed at finding previously unknown phenomena in the data through prediction and description, while verification-oriented goals are aimed at verifying an existing or known phenomenon that is implied in the data through the description and hypothesis testing. Examples of data mining tasks may include Regression, classification, rules induction, segmentation, association, dimensionality reduction, hypothesis testing, correlation and feature extraction and analysis [42]–[44].

Data mining employs a wide spectrum of techniques in order to achieve its goals and perform its tasks which are adopted from fields, including machine learning, statistics and pattern recognition. Data mining techniques can be either

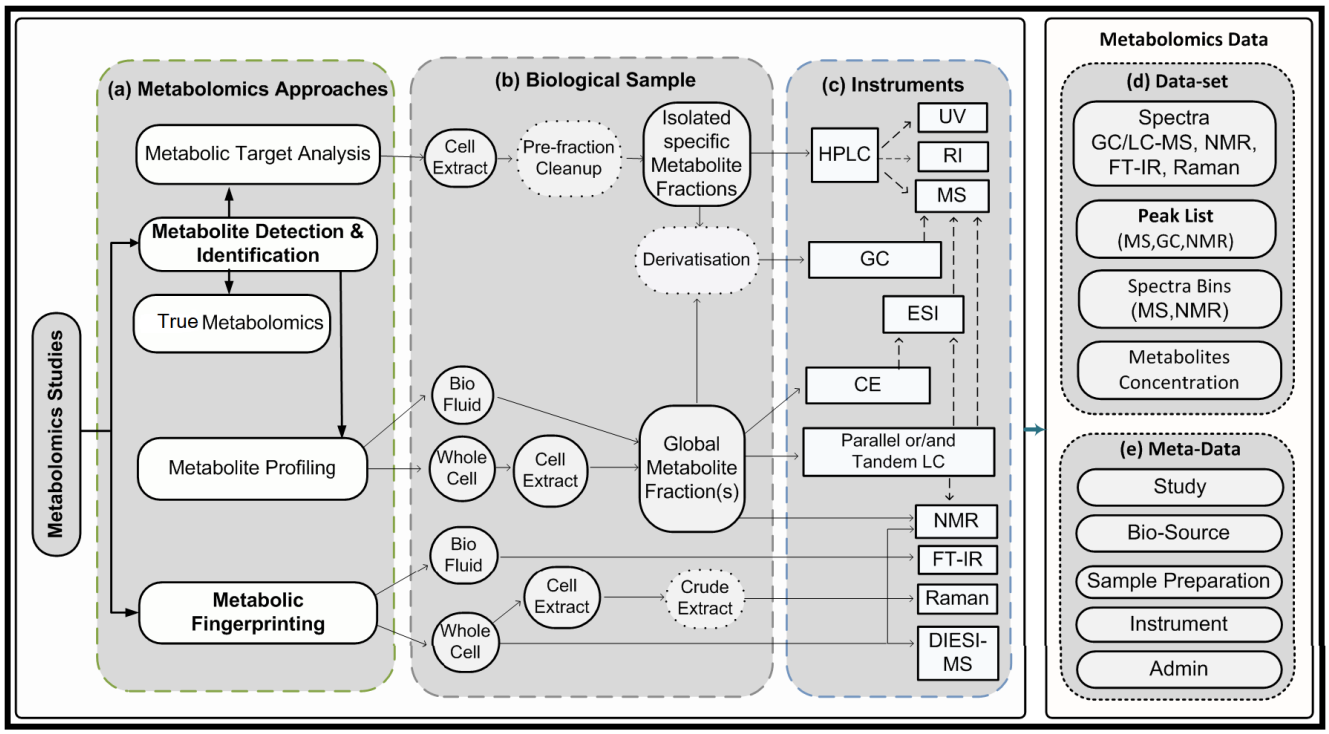

FIGURE 1. **The Design of Metabolomics Study- A diagram illustrating the relationship between metabolic approaches. It also shows the relationship between metabolic approaches and bio-samples, and between sample preparation and data acquisition instruments.**

supervised or unsupervised. While unsupervised techniques allow learning from data through finding patterns or natural grouping with no guidance in unlabeled data supervised techniques are used to learn with guidance from labeled data [22]. The mechanism of mapping the aims of metabolomics investigations and studies to the data mining goals, tasks and techniques are discussed in [45] which proposes a strategy for selecting data mining techniques based on this mechanism as well as based on the nature of the data. It also offers demonstrative examples for the mapping mechanism and the selection strategy using a number of metabolomics applications.

A data mining process model would help provide a systematic approach and a formalised framework for analysing metabolomics data. A Data mining process can provide a systematic approach and formalised framework for analysing metabolomics data. This can be done through applying a well-defined process which has clear guidance and implements well-organised and clearly described steps. However, it is argued that existing process models suffer from major flaws in their design and in their poor support for practical issues. In addition, the generic nature of most of the existing data mining process models and their bias towards business, makes them unsuitable for performing data mining in science. This suggests the need for a new data mining process model that must consider the scientific nature of metabolomics data mining and provides support to both inductive and deductive knowledge discovery which can be performed either data-driven or hypothesis-driven data mining [41], [42].

III. METHODS

The design of the proposed process model (MeKDDaM) went through several development stages. Figure 2 illustrates the stages of the research methodology that was conducted.

In the first stage, the requirements of the process model were identified based on the fundamentals of metabolomics data analysis as discussed in Section II-A. The requirements covered the nature and formats of metabolomics data and its required preprocessing, pretreatment, data analysis and interpretation procedures. It also covered the consistency with the cycle of knowledge and the satisfaction of other scientific requirements including traceability, justifiability and the reproduction of the data analysis results. The requirements also considered the practical aspects of the data mining process in respect to its manageability, human interaction, quality assurance, and both metabolomics and data mining standards and a number of other data mining desirable features including visualisation, data exploration, knowledge presentation, and automation.

In the second stage, existing data mining process models were revisited and critically analysed in terms of their structure, flow, shortcomings, and good features with respect to satisfaction of the identified requirements, and compliance with the principles and good practices of process engineering. This stage also involved inspiration from the principles, best practices, and recommendations of a number of relevant interdisciplinary fields including Process Engineering, Software Engineering, Machine Learning and Scientific Methodology aimed at satisfying the requirement of metabolomics data

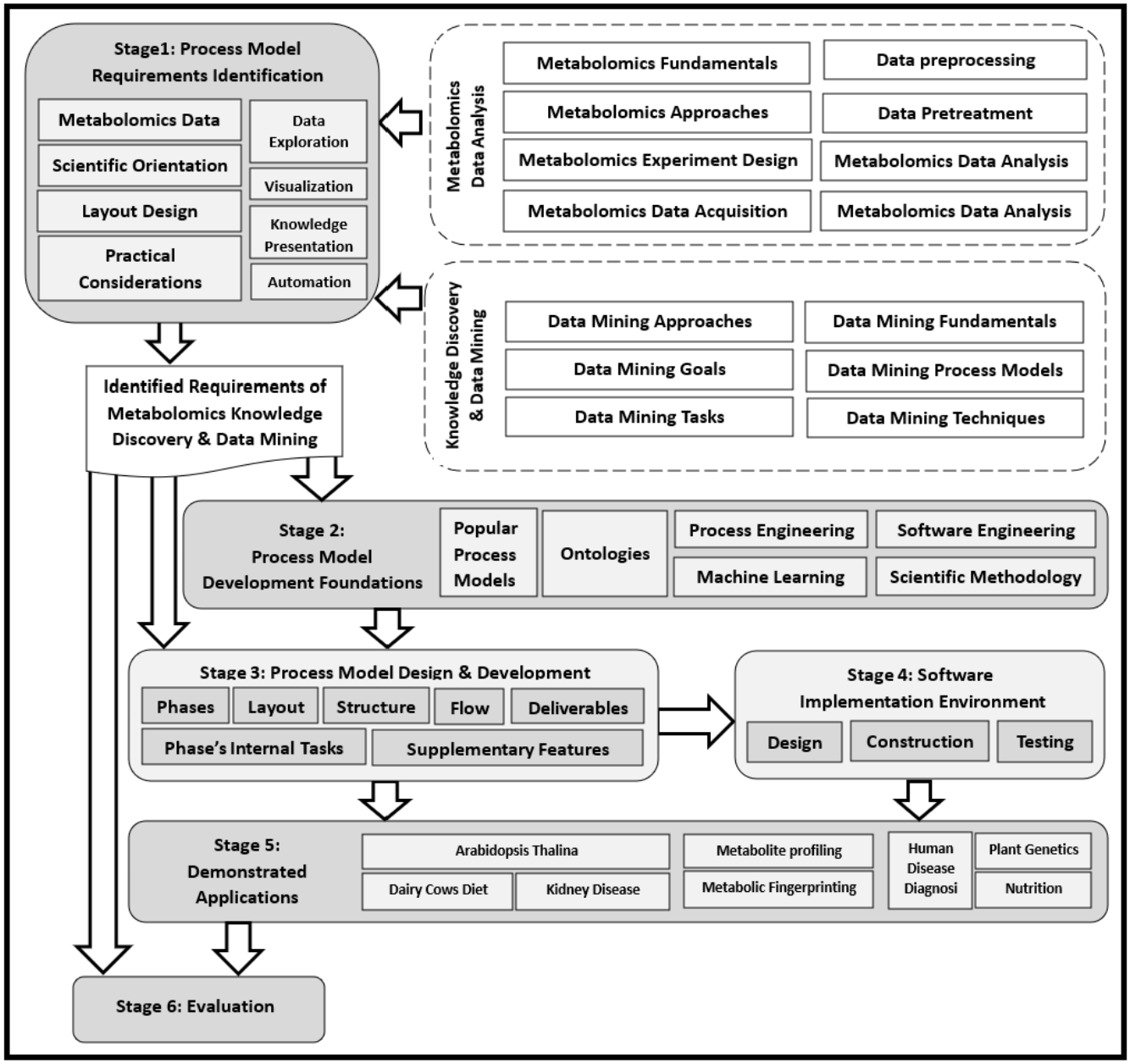

FIGURE 2. **The Research Methodology- The six stages of the research methodology conducted to develop MeKDDaM.**

mining; enhancing the design of the proposed model; and addressing the gaps, weaknesses and shortcomings of existing process models. This forms the foundations of the design of proposed process model.

In the third stage, the proposed process model was developed based on the identified requirements, the existing process models and the inspirations offered by the fields considered. The design of the process model started with identifying the process phases and justifying their context, rational and scope based on the requirements of metabolomics data mining. The identified process phases were then organised and the process model structure and flow were laid out to create a prototype that was based on the process model development foundations and some features that were already found in existing data mining process models. The prototype was improved through several iterations, validations, and verification and was then described thoroughly using a graphical representation and accompanied with possible execution scenarios.

In the fourth stage, MeKDDaM was realised using a software environment that was named MeKDDaM-SAGA [15], [16]. The software environment provides a realisation of the process structure and flow. The software executes the process execution externally or internally using a number

of embedded facilities that allows the execution of a number of internal automated tasks and activities such as: data preprocessing, data exploration, data acclimatisation, model building, model evaluation and visualisation.

In the fifth stage, MeKDDaM was then applied to four real-world metabolomics applications aimed at confirming the process model's satisfaction of the requirements of metabolomics data mining and highlight its strengths and unique features. The applications covered two plant genetics (***Arabidopsis thaliana***) studies, an animal nutrition (***Dairy Cows Diet***) and a human disease (***Kidney Disease***) study [14]. These applications were carried out to demonstrate the applicability of the process to (1) various metabolomics investigations, approaches, and instruments, 92) different data mining approaches, goals, tasks and techniques (3) both data-driven and hypothesis-driven data mining.

In the sixth stage, MeKDDaM was evaluated in terms of its satisfaction of the identified metabolomics requirements, scientific orientation and the principles of its layout design. In addition, the process model was also evaluated in terms of its consideration of the practical aspects including management, human interaction, quality and standards as well as its support for a number of desirable features including visualisation, **data exploration**, **knowledge presentation** and automation.

A. REQUIREMENTS IDENTIFICATION

Identifying the requirements of metabolomics data analysis and mining was the backbone of developing the process model. It covered the requirements of metabolomics data, scientific orientation, layout design in addition to process practical considerations including management, human involvement, quality assurance and standardisation. It also covers a number of other desired features including data exploration, visualisation and automation.

1) METABOLOMICS DATA

Metabolomics data consists of both the dataset that is generated by the data acquisition instrument and its associated meta-data. The dataset varies depending on the data acquisition technique [30]. Here we describe the data generated by some of these instruments. DIMS data can be represented through 2D matrix $m\,X\,f$, where m is the unified length of the vector, while f is the number of files. In the case of handling multiple scans, a third dimension can be added to the matrix to represent the retention time [20]. MS based techniques data is represented in a table that stores mass-to-charge-ratio *(m/z)*, retention time channels and their corresponding intensities. The selected peaks are integrated at a specific *m/z* channel and used in the peak table. The mass spectra of Gas chromatography-Mass spectrometry (GC-MS) can be represented as a bar plot of normalised ion abundance versus m/z. GC-MS multiple scans data can be stored using a 3D matrix, where the number of scans is associated with retention time. High-performance liquid chromatography (HPLC-MS) data is similar to liquid chromatography-mass spectrometry(LC-MS) and GC-MS data in terms of its structure and format, however the qualitative characteristics of these data are different as well its meaning. Nuclear Magnetic Resonance (NMR) data takes the form of peaks in a spectrum which represent metabolites signals. These peaks are usually stored in a table in terms of ppm channels and their corresponding intensities and they can be then integrated at a specific channel to form a peak table [46]. Fourier-Transform Infrared Spectroscopy (FT-IR) spectrometry data are generated as an interferogram which is usually transformed using Fourier Transformation in order to decode the individual frequencies and produce the final spectrum [47] which is usually done in order to take into account the combined performance of source, interferometer and detector. On the other hand, meta-data consists of other data that are collected regarding metabolomics study and might influence the dataset such as bio-source, sample preparation, metabolic approach, data acquisition instruments in addition to other administrative data that may be related to the study [31], [46], [48], [49]. The proposed process model must consider the nature of metabolomics data and allow the performance of effective data preprocessing and pretreatment procedures in order to make sense of the data and convert it into knowledge.

Data preprocessing as defined by the Metabolomics Standards Initiative (MSI) [50], [51] refers to the set of procedures that are performed in order to avoid error propagation and other issues related to the assay and data acquisition instruments. Examples of preprocessing procedures may include: data cleaning and transformation such as deconvolution, peak alignment, peak labelling, profile alignment, spectral transformations, binning, peak listing, concentration profiling, and other procedures such as normalisation in reference to internal or external standards and handling the missing values related to the machine detection limit [46], [52] or those related to the design of metabolomics study that may influence the data analysis final result e.g. limitation of machine detection and the use of internal or external standards [28], [30]. Data preprocessing procedures must be comprehensive and thorough. However, the scale and intensity of the data preprocessing procedures depends on the experimental protocols, conditions, sampling and sample preparation procedures [52] as well as on the particular data acquisition technique and instrument and on how much preprocessing was performed using the instrument embedded software.

The MSI defines pretreatment as the transformation procedures which concern the nature of the data. Data pretreatment aims to clean and transform the data for further data mining and analysis [46]. However, the particular pretreatment procedures depend on the needs of the data mining techniques which are used in model building, as well as the statistical characteristics of the acquired data set. Pretreatment procedures may include a wide range of methods, e.g. mean-centering, auto-scaling, range scaling, missing data and outlier handling in addition to power and log transformations [30], [53], [54].

2) SCIENTIFIC ORIENTATION

The scientific nature of metabolomics investigations must be considered in the design of the process model. It must be consistent with the metabolomics cycle of knowledge and it should pay attention to issues such hypothesis formulation and testing, decisions justification, traceability of process model procedures and reproducibility of its generated result. In addition, the process model must support both hypothesis-driven and data-driven data mining approaches. The process decisions and procedures must be justified and reported to ensure their validity and allow the traceability of their outcomes which contributes toward the overall reproducibility of the metabolomics analysis.

Metabolomics is consistent with the cycle of knowledge [1] which supports both deductive and inductive knowledge acquisition. Hypothesis generation and testing are important aspects in the cycle of knowledge. The generation of observation-based knowledge is conducted either via induction which aims to generate rules or models based on observations or via deduction based on testing hypotheses. However, in both cases, hypothesis generation is important in the design of metabolomics study where observation plays a significant role in building predictive metabolomics models and for the purpose of generating and testing [25], [55]. On the other hand, data mining is viewed by many as consistent with both hypothesis-driven and data-driven data mining approaches of scientific methodology (observation-hypothesis-experiment). Since data mining generates hypothetical results rather than facts, this assumption does not affect its validity as a scientific tool, particularly with experimental studies which depends on hypothesis generating and testing such as metabolomics [56].

3) LAYOUT DESIGN

The process model must be organised in a structured, understandable and well-defined fashion. However, simplicity is also an important and desirable requirement for the design of any process model. The process model must be laid out in an understandable, well-structured and well-organised fashion. The process model must define the process flow and iteration, as well as the relationships between its phases, either with their successor phases or in terms of feedback to their predecessors. Reducing the complexity of the process model is an important aspect which must be considered in the design of the proposed process model. Understandability is an important requirement for an effective data mining process. It enhances the applicability of the process in the real-world and supports the validity of its execution.

Coherence of the process model phases is also an important aspect in the design of the proposed process model. Data mining process in general, and metabolomics data mining process in particular, involves a huge number of activities, which must be organised in a structural and well-defined fashion. Coherence of the tasks within the process phases reduces its complexity and enhances the modularity of its phases, which is an important principle in process engineering. The unity and independence of the process phases, as well as their tasks integration reduces the overlapping between the process phases, simplifying the relationships between these phases.

Iteration organises the repetition of one or more of its activities in a loop-like fashion. Iteration can be seen as a self-contained mini-project that is made of a set of activities, which are repeated for a particular number of times [57]. The iterative nature of Metabolomics data analysis was neglected by most researchers in the metabolomics community. This is quite apparent in viewing metabolomics either as a pipeline [58] or as a workflow of sequential activities [59]. On the other hand, the iterative nature of data mining process was emphasised early in data mining [60]. Most of the existing process models emphasise the iterative nature of the data mining process, which is realised as feedback between one or more of the process phases, or as iterative execution of the entire process as a cyclic loop which is terminated depending on a defined exit condition.

In order for the process model to be generalisable to cover all types of metabolomics studies and also to be customisable to be applicable to the particular metabolomics study, the design of process model must utilise the concept of mapping which is available in CRISP-DM. CRISP-DM consists of four models each containing a set of tasks which move from generality in the top level to specificity in the lower levels. The phases in the top level are meant to be as general as possible to be applicable to all applications. The tasks are stable enough to be valid for new mining techniques that may evolve over time. Generic tasks in the second level are linked to specialised tasks in the third level which describes the applicable actions that suit a specific application in practise. The fourth level represent instances of specialised tasks in level three where tasks can be transformed into actions, decision and results which describe specific practical implementation [61], [62]. This concept allows covering all metabolomics data mining applications, while satisfying the special needs of a particular metabolomics study. It describes a number of generic tasks that are customised to a more specific ones, which are more relevant to the particular metabolomics study and specific domain of the process model application [63].

4) PRACTICAL CONSIDERATIONS

The real-world applications of metabolomics data mining is a complex process. It require considerable time, cost, and expertise and it involves a large number of intensive data processing and modelling procedures [60]. The practical aspects of the process must also be considered including project management, human interaction, quality assurance and standardisation.

a: PROJECT MANAGEMENT

In order to tackle the practical complexity of metabolomics data mining, effective project management must be incorporated in every aspect of the process model [63]. It should

define a framework for project planning, management, monitoring and control and it should cover task breakdown and organisation, resources management and allocation, feasibility assessment and estimation, success definition and measurability [64]–[66].

b: HUMAN INTERACTION

is important for the success of data mining [60], [67]–[69]. Organising human expert interaction with the process helps provide guidance regarding the activities to be performed and the decisions to be made throughout the process [65], [67], [68], [70]. Human interaction is important particularly in objective definition, results interpretation and validations which may require the use of automated tools for both data and results visualisation, as well as in results comparison and validation [71]. The main players in the process model are the domain expert and data miner where each of them complements the other's role on both the technical and contextual levels. [65], [67], [68].

c: STANDARDS SUPPORT

is encouraged by both data mining and metabolomics [51], [63]. Data mining and metabolomics standards must be considered in the design of the proposed process model and also in reporting the results of its execution. Metabolomics standards focus on metabolomics reporting structures, while data mining standards concerns the process modelling procedures and the format of its deliveries. Data mining standards enhance the process manageability, promote the achievements of its analytical objectives and ensure the validity of its results. They also help controlling the process vocabularies and formalising its structure, procedures and deliveries [63] such as XML and PMML [65], [72] which are recommended by the Data Mining Initiative [65], [73]. In addition, metabolomics standards must also be used in defining the process model procedures and deliverables including MSI [50], [51], ArMet [74], MeMo [48], RSBI [33], ISA-TAB [33], and FuGE [32].

d: QUALITY ASSURANCE

concerns the quality of data and results and the validity of the analytical procedures applied. The quality of the process is a key for ensuring the quality of data mining results and for justifying and ensuring the validity of its discovered knowledge and underlying models [58]. Validating the quality of data is crucial for the success of data mining. Poor data quality may hide useful patterns in data, may hide interesting phenomena, and may lead to false discoveries. Outliers and missing values might affect the results of data mining and may lead to false knowledge discovery [60]. Noise and errors might enter the data during any stage in metabolomics assay including sampling, sample processing and preparation, as well as during data acquisition [75]. The process model must validate the execution of the process and provide mechanisms for validating the activities performed within the process phases and their deliveries.

5) OTHER DESIRABLE FEATURES

A number of other desired features were also identified. These were based on recommendations reported in metabolomics and data mining literature in addition to other useful features found in some of the existing data mining process models. They include: Data Exploration; Knowledge Presentation, Visualisation and Automation. These features were then used in the design of the process layout and for identifying its phases.

a: DATA EXPLORATION

covers data understanding, investigation and prospecting. Data understanding involves verifying the understandability of data through describing the meaning of the data variables, and the scope of their values and the characterisation of its structures, formats, data types [76], [77], volume and the ratio between the number of the its attributes to the number of its instances [78]. Data investigation covers issues which concern the quality of metabolomics data such as missing values [79], [80] and outliers [81]. Data prospecting aims to prospect the potential of data in terms of its trends, distribution, tendencies [76] and relationships and to confirm the data relevancy, sufficiency and adequacy for achieving the data mining objectives. Data mining techniques may vary in their sensitivity, tolerance, and response to issues such as the existence of missing values, outliers, the distortion of the data distribution [79]–[81] and the sufficiency of data observations [78]. Some data mining techniques are able to handle some data types rather than others, which in this case may require further conversions. Some aspects of data exploration are available in some existing data mining process models e.g. CRISP-DM [61], EBPM [82], Two-Crows [83].

b: KNOWLEDGE PRESENTATION

provides effective mechanisms for converting the model into knowledge. Brachman and Anand (1994) emphasise the importance of knowledge presentation in the data mining process, while Hall (2006) suggests that knowledge presentation contributes to understandability, which is crucial for the success of data mining in metabolomics. In addition, knowledge presentation is important for human interaction as it helps the domain expert in the interpretation and evaluation of the acquired knowledge.

c: VISUALISATION

is an important feature in the data mining process in general and in metabolomics data mining in particular. Visualisation facilities are required throughout the various phases of the data mining process and it is used in order to attain human interaction and enable other confirmatory involvements [84], [85]. Visualisation enables the comprehension of knowledge and gaining insight into it [84], [86]. Most visualisation techniques used in metabolomics are based on multidimensional scaling and projection [87]. Examples of other techniques used for visualisation in metabolomics include: clustering

diagrams, dendrograms, network diagrams, and heat maps [22], [87].

d: AUTOMATION

is a desirable feature in metabolomics data mining. While full automation of the data mining process is mostly unlikely due to the need for human judgement in some of its process phases such as objective definition and knowledge evaluation. However, partial automation can still be achieved in phases such as data exploration, data acclimatisation, model building and model evaluation [69], [85], [88]. The importance of the automation for data mining attracted the attention of practitioners from the beginning [60] as the need for an integrated environment was emphasised in order to enable the users to apply the complex data mining process. However, until now, no serious efforts have been made towards addressing the automation of the data mining process.

B. PROCESS MODEL DEVELOPMENT FOUNDATIONS

The process model development foundations aim at identifying the potential solutions that satisfy the requirement of metabolomics data mining including the nature, quality and format of its data and its involved preprocessing, pretreatment, data analysis and interpretation procedures in addition to the requirements of the process design that are related to the process model layout design, scientific orientation, practical considerations (project management, standards support, human interaction and quality assurance) and other desirable features including data exploration, knowledge presentation, visualisation, and automation support.

The process development foundations and proposed solutions are based on the ontologies of scientific experiments, metabolomics, and data mining that are used to confirm the process validity and to control its involved vocabulary on one hand, and on the principles, best practices and inspiration that have been adopted from a number of cross-dictionary fields including Process Engineering, Software Engineering, Machine Learning and Scientific Methodology on the other. These inspired foundations are used to provide several enhancements and improvements which have been injected into the development and design of the proposed process model in order to satisfy the requirement of the process model development identified in Section III-A and to address the shortcomings of the existing data mining process models discussed earlier in Section III-B1. Figure 4 illustrates the fields inspiring the process development. Figure 3 illustrates the relationship between the processes and process models of data mining, software engineering, project management, and metabolomics on one hand, and the process meta-model, on the other hand.

1) POPULAR DATA MINING PROCESS MODELS

A critical analysis was carried out which covered ten of the popular data mining process models including KDD [39], [60], FS-FE [91], 5As [92], CRISP-DM [61], Two-Crows [83], EBPM [82], Nautilus [93], SEMMA [94], Kantardzic [86] and Rapid Prototyping [95]. The process models were evaluated on their satisfaction of metabolomics data mining requirements which covered: scientific orientation, practical considerations and the desired process features as discussed earlier. The critical analysis identified a number of gaps, weaknesses and shortcomings to be addressed and a number of good features to be utilised in the design of the proposed process models. The review revealed the lack of scientific orientation in most of these models and the poor support of management, human interaction, standardisation and quality assurance in most of the process models. It also revealed the inadequate support for visualisation, data exploration, knowledge presentation and automation. Table 1 scores the existing data mining process models performance against metabolomics data mining identified desirable requirements. These conclusions were confirmed by later studies which agreed with the results of our critical analysis [3]–[8] and suggested addressing some of these issues either through extending the existing process models [9], [10], or through proposing new generic [11] and dedicated process models [12], [13].

2) PROCESS ENGINEERING

A process model is a classification of processes of the same nature into a model, which describe the process on the type level. A process model is used for developing various applications, which act as instances of the model. A process model provides guidelines for how things should or could be conducted, while the process is what really happens [96].

Process engineering principles are important for the design of data mining process models. The focus of process engineering is providing a formalised and conceptual framework of the engineering process. A process model can provide a framework and formalised guidelines for conducting the engineering process in a systematic fashion. This would help the validity of the procedures involved, as well as assuring the quality of their outcomes. Figure 5 illustrates the relationship between the processes and process models of data mining, software engineering, project management, and metabolomics on one hand, and the process meta-model, on the other hand.

The principles of process engineering in general and in software engineering in particular have been discussed in [96]. Below is a summary of some of the important aspects of these principles:

- Abstraction: refers to the type level of the process model, where it provides a classification of processes of similar type.
- Notation: concerns how the process model is represented and described, as well as the semantics of its representation, e.g. natural language, diagrams.
- Content: refers to the coverage and granularity of the process model where coverage refers to the orientation of the process in terms of activity, decision and product orientation, while granularity refers to the level of

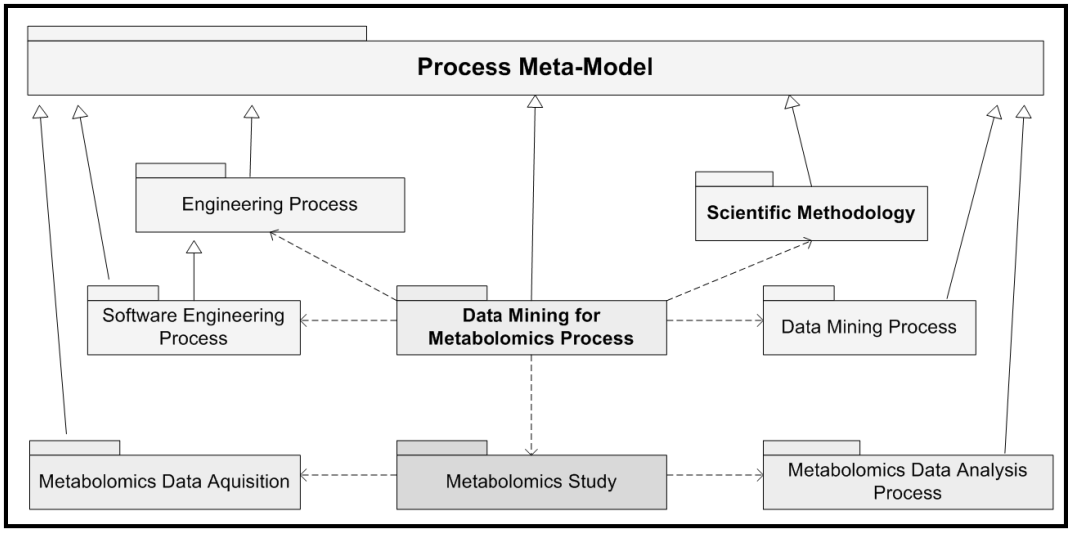

FIGURE 3. **Process models Meta Models:- A UML representation of the relationship between the process models of data mining, software engineering, project management, and metabolomics using UML generalisation notation which depicts is-a relationships.**

TABLE 1. **Scoring Matrix: A six-categories scoring for the process models satisfaction of metabolomics data mining desirable requirements. Five pluses for complete satisfaction, four pluses considerable satisfaction, three pluses acceptable satisfaction, two pluses some satisfaction, one minimal satisfaction, and no pluses for no satisfaction.**

Desirable Requirement	KDD	EBMP	Prototyping	Two-rows	CRISP-DM	Nautilus	Kantardzic	5As	SEMMA	FS-FE
Scientific Orientation	+++	+++	++++	++	++	++	++++	+	+	+++
Layout Design	+++	+++	++++	+++	++++	+++	+++	+++	+++	+++
Management Support	+++	+++	++++	++	+++	++	++	+	++	+++
Human Interaction	++	++	++++	++	+++	++	++	+	+	++
Quality Assurance	++	++++	++++	+++	++++	+++	++	+	++	++
Standards Support	+++	++++	+++	++	+++	++	++	+	+	+++
Visualisation Support	++	+	+	+	+++	+++	+	+	+	+
Data Exploration Support	+	++++	+++	++++	++	+	+	+++	++++	+
Knowledge Presentation Support	+++	++	+++	++	++	++	+++	++	+	+++
Automation Support	++	+++	++	+++	+++	++++	++	++++	+	+

detail represented and provided by the process model, e.g. large, fine and variable where large represents the highest level of the process model description, fine represents the details of the process model, while variable represents the customisable level of the model.

- Modularisation: concerns the unity of the parts of the model, as well as their level of abstraction and aggregation. The former is linked to the granularity attribute of the process models abstraction, while the latter concerns the relationship between different parts of the model
- Instantiation: refers to the application of the process model to a particular situation by creating an instance process of the generic process model.

[97] suggests a number of steps, which involves the development of a process model, which include technology provision, process requirements analysis, process design, process implementation, and process assessment. The development of the proposed process model is consistent with these steps. However, due to the nature of the domain application, which involves both metabolomics and data mining, some of these steps are split into two or more.

3) SOFTWARE ENGINEERING

Some argue that "Knowledge discovery should follow the example of other engineering disciplines that already have established models. A good example is the software engineering field" [63]. Software engineering has many characteristics which are logically relevant to knowledge discovery and data mining. Figure 5 suggests a possible mapping between data mining process and the software engineering process depending on their shared engineering approach. It illustrates the relationships between a number of common data mining phases, which are explicitly described by the majority of the existing data mining process models and their equivalent phases in the software engineering process. The generic engineering process in this analogy acts as a bridge between the two.

Software engineering adopted its systematic engineering approach from other engineering disciplines which have achieved considerable success across the known history of humanity from the construction of pyramids, to the building of the gardens of Babylon, to the industrial revolution and today's constructs and machinery's. Software engineering

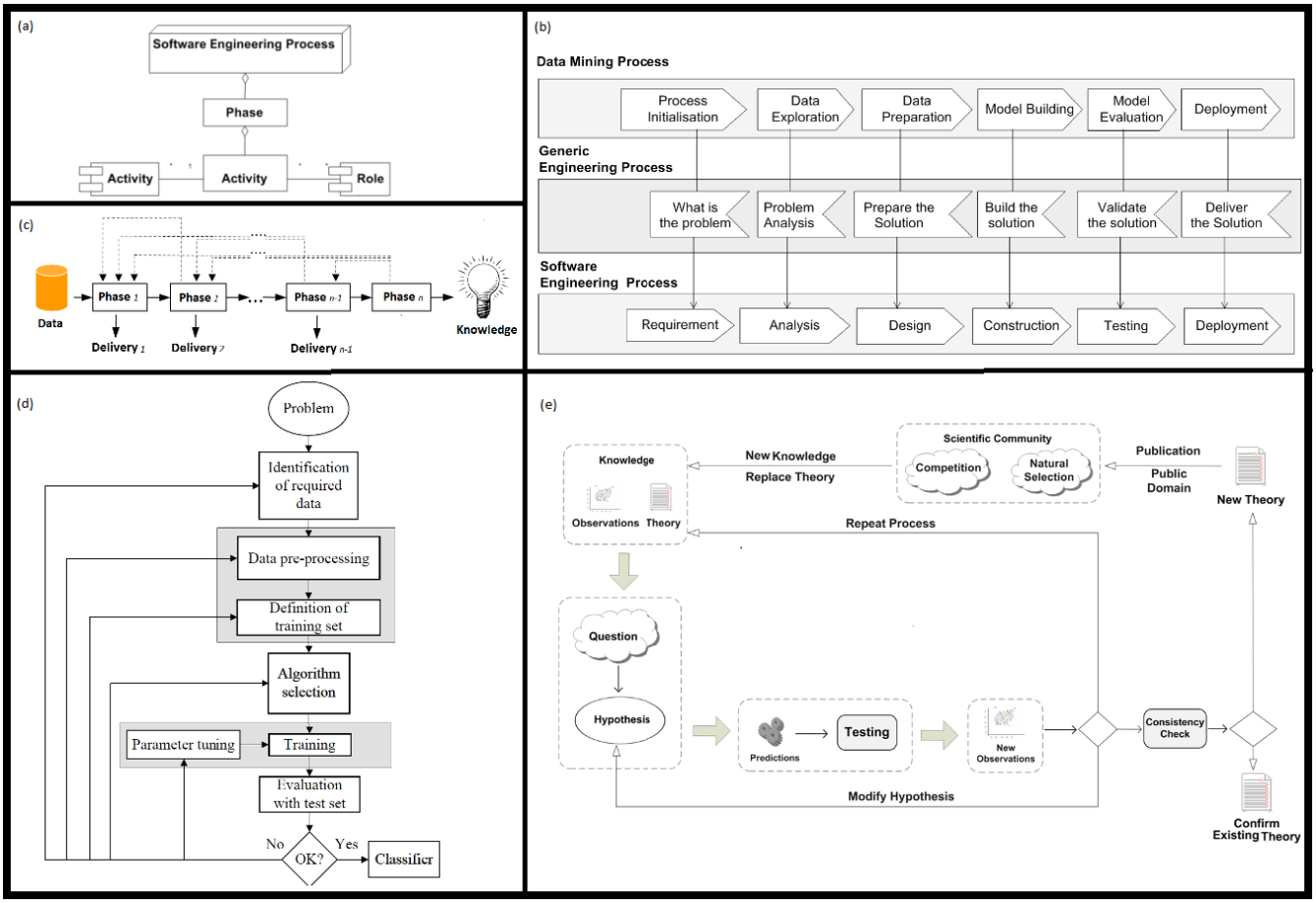

FIGURE 4. **Process model design inspirations:- (a) Illustrates a software engineering process meta-model [89], a demonstration using UML notation. diamond head arrows represents part of relationship while socket-like-rectangle represents activity components (b) Shows the mapping between the Software engineering process, data mining process and generic engineering process. (c) Illustrates the general structure of a knowledge discovery and data mining process (d) illustrates using a flow-chart machine learning classification process [90](e) Illustrates the process of conducting hypothesis-driven inquiries based on the scientific methodology.**

process models or life cycles have been in use for several decades. A review of process software engineering processes is available in [89], [98]–[100]. Below we provide a summary of these process models.

- Waterfall: known for its simplicity and understandability. The model has well structured phases and well-defined deliveries.
- Iterative Waterfall: a iterative variation of the waterfall model where feedback is allowed to previous and earlier phases.
- Incremental Model: encourages prioritisation of objectives, by doing the important ones first and moves to the less important ones. It also provides a framework for accumulative development as a snowball.
- Spiral Model: focuses on involving domain experts in each of its spiral iterations, as well as incorporating an accumulative and continuous evaluation.
- Prototyping Model: focuses on creating a working prototype model from the earlier stages and emphasises the role of domain expert in the process. The prototype is enhanced iteratively until domain expert satisfaction is gained.
- V-model: organises the process phases in a V-like shape which provided an extra level of report-based quality assurance, which is known as Cross-delivery validation. The earlier deliveries in the process are validated in terms of their correctness, completeness, and consistency with those generated in later phases. [99], [101]–[103].
- Rational Unified Process: provides two dimensions for the process execution that are called workflow and phases. It also adds three other extra activities that supports the execution of other core activities in process workflow: (1) project management; (2) configuration management; and (3) environment. The process model considers the overlapping nature of the process phases and facilitates the automation of process activities execution.
- Agile Model: encourages more human interaction between the development team and also with the domain expert. It also provides flexibility to the development

process by allowing changes to occur at any stage.

The utilisation of software engineering in the design of data mining process models has attracted attention from the inception of data mining. This is noticeable in KDD and CRISP-DM, as well as in other data mining process models. In addition, this argument was also supported by [63] and [65]. Due to the similarity between software engineering process and data mining process, the engineering approach of software engineering methodologies can be used to tackle the challenges in data mining and enhances its process. The relatively longer history of software engineering compared to data mining, has allowed it to achieve a relative maturity and success in tackling a range of challenges regarding the complex development process including the structure, manageability, validity and quality of its process, as well as its support for human interaction, quality and standards which will be discussed later in the section. [63] argues that knowledge discovery and the data mining process should take advantage of the successful approaches that are used in software engineering.

The principles and best practices of software engineering have been considered in the development of the proposed process model. It has been utilised in the process model layout design and flow, iteration and feedback, as well as in the selection of its phases, which will be discussed in Section III-C. [89] introduces a meta-model for describing the software engineering process, which is illustrated in Figure 5.

4) MACHINE LEARNING

Machine learning is defined as "*the study of computational methods for improving performance by mechanising the acquisition of knowledge from experience*" [104]. Machine learning achieves its learning goals by finding regularities in data, and improves its performance and the validity of its learnt knowledge by using an independent set of data. Machine learning aims at building a model which is based on existing data instances in order to generate a generalised hypothesis that can be used for predicting future instances using predictor features [90]. Machine learning uses a wide range of supervised and unsupervised techniques in order to learn knowledge from data.

It was crucial for the design of the new data mining process to consider those techniques, as well as the process of machine learning illustrated using flow charts in Figure 5. It can be argued that in the case of the data mining process, the stage of "*Algorithm Selection*" must be extended to cover other machine techniques which are available for data mining. In this case, technique selection must be moved to precede both " *Data Preprocessing*" and "*Definition of Training Set*". Yet, those two stages can be merged into one stage which can be called data acclimatisation as they both aim to make the data suitable for modelling by the selected technique. It may be worth noting here that data preprocessing in this context is similar to the data preparation phase in most data mining process models and it must not be confused with metabolomics data preprocessing which has already been discussed earlier. On the other hand, the "*Training*" stage must be renamed as model building. Training is only one aspect of model building, which also covers deriving a model and generalising as defined machine learning. The earlier stages of the process illustrated seem to fall out of the scope of machine learning, as it is already covered by earlier data mining phases. Therefore, the data mining process must include phases which should be allocated for carrying out important machine learning stages including technique selection, data acclimatisation, model building, and model testing.

5) SCIENTIFIC METHODOLOGY

The Scientific Method is defined as the logical scheme which is applied to answer a particular scientific question [105] and suggested explanation based on observation and followed by controlled experiments designed and executed in order to validate, refine, or reject a pre-formulated hypothesis [106]. Scientific methodology was pioneered by the medieval Arabic scholar al-Hasan Ibn al-Haytham, who is also known in the west as Alhazen [107]. It was then developed by Galileo Galilei and later by other thinkers including Bacon, Descartes and Pierce [106]. Figure 5 illustrates the process of performing hypothesis driven inquiries based on scientific methodology. Therefore, scientific methodology can also be utilised in order to improve data mining processes and to enhance the design of the process model. It can be argued that the data mining process in general is in fact consistent with scientific methodology, particularly when it comes to hypothesis driven data mining where the existing observation and theory can be mapped to the targeted data, while the question and perhaps some aspect of hypothesis formulation can be mapped to objectives definition, the answer deduction and observation prediction in this case can be mapped to model building and training. Testing can serve the same purpose in both processes and therefore, validation can be adopted to be embedded in every aspect of the proposed process model, either on the level of the process or on the level of phases, and it must cover both process data and procedures.

Justification and tractability are also essential for scientific methodology. These are crucial for the consistency check of the generated results (results reproducibility), as well as for reasoning. Reasoning is an important issue in scientific experiments where it seeks to explain observations and phenomena. However, reasoning is not always possible in metabolomics. This is due to the limitation of human knowledge and due to the complexity of the observed metabolic phenomena. The proposed model must incorporate justification and tractability in all of its features on the level of its procedures, as well as on the level of decisions and deliveries. It must also encourage reasoning where possible.

6) ONTOLOGIES

An ontology is a content theory describing the classes of objects, their properties, and the relationship among them in a specified domain of knowledge. It defines a common and controlled vocabulary in the domain application, as well as describing things as they are, categorising them, and connecting them to each other and to their context in their domain [108], [109]. Ontologies are useful in the development of the proposed process model in order to confirm the understanding of the concepts involved in metabolomics and data mining and to control their vocabularies as well as to validate the process model structure, and the correctness and completeness of its involved procedures. Ontologies provide means for capturing, categorising, and describing the concepts of both metabolomics and data mining, and verify their comprehension. They also encourage the use of a controlled vocabulary, which can be used for describing the process model and facilitating the analysis of the requirements. A knowledge discovery and data mining ontology provides unified and formalised means for describing it.

On one hand, and despite the fact that there have been recently several attempts to propose a unified ontology for knowledge discovery and data mining, each of these focused on a different aspect of data mining and knowledge discovery. OntoDM covers both data mining procedural aspects and other aspects related to its techniques [110], while KDDONTO was proposed in [111] which describes the procedures involved in knowledge discovery and data mining process as well as their algorithms and data. Bernstein's proposed an ontology that perceives data mining as a knowledge centred process and focuses on the knowledge discovery aspect of the data mining process while it also describes its tasks and algorithms [112]. Other efforts in data mining ontology are proposed in [113] which focuses on the automation aspects of the data mining process and in [114] which focuses on the planning and management aspects of the process.

On the other hand, less work has been dedicated to metabolomics ontology. Despite the fact that metabolomics ontology is often discussed in the context of metabolomics reporting standards as in [23], [51], [74], [115], and a metabolomics ontology was considered in the broader picture of functional genomics ontology including FuGe [32] and OBO [32] and MeMo [48], MeltDB is the only standard which provides discussion regarding metabolomics experiment ontology, but this is only done in the context of its proposed software platform [116] which constrains its coverage. RSBI (ISA-TAB) provides a considerable effort towards controlling metabolomics investigations vocabulary, which is adopted in this work for the analysis and description of the proposed process. The MSI is currently developing an ontology for metabolomics through its Ontology Work Group (OWG) [59]. Scientific Experiments Ontology(EXPO) was proposed [117], which was designed as a universal ontology in order to generalise the subject specific standards and ontologies that are available in bioinformatics, e.g. MeMo, FuGO, MSI and SUMO [118].

C. PROCESS MODEL DEVELOPMENT

Here we discuss the evolution of the process model layout design across three major versions. Although other minor changes were made during the process development, they have been merged with these three versions. The evolution of the process model design has been driven by the elaboration of the requirements discussed in Section III-A, as well as by the ideas inspired by the development foundations discussed in Section III-B including the existing data mining process models, scientific methodology, process engineering principles, software engineering methodology, machine learning, and their relevant ontologies. Figure 5 shows a UML diagram that demonstrates the foundations of the process model development and its relationship with the design of the process model. We also provide a discussion of the rationale and justification for the selection of the process model phases. It also discusses various concepts, mechanisms and procedures which are relevant to the scope and context of the process phases based on the literature and in the light of the process development requirements and foundations discussed in Section III-B.

1) PHASES IDENTIFICATION

Metabolomics data mining, involves performing a large number of activities which must be organised into a set of self-contained phases which must be designed in an abstract, logical and cohesive manner [63] in order to increase the unity and independence of each phase and maximise the integration among their internal tasks while reducing the overlap between phases and simplifies their relationships, in addition to facilitating a better management and validation of the activities within these phases. The process phases were identified based on metabolomics data analysis workflow as reported by the MSI and based on the phases which were found in common in current popular data mining process models. The ontologies of data mining e.g. OntoDM [110], KDDONTO [111] and KDProcess ontologies were used in the design of the process model [112], while EXPO [117], FuGe [32] MeMO [48] and MSI [50], [51] were used to confirm the domain concepts understandability, correctness and coverage and to control the process vocabulary required. The scope of the process phases and their internal structure and subtasks were inspired by use-case description in software engineering and was enhanced based on the principles of project management in addition to some good features available in the surveyed data mining process models. Each phase is described by its subtasks including their prerequisites, objectives, participants, and deliveries. Figure 6 illustrates the relationship between the process model requirements and development foundations and phases identification.

The Objectives Definition Phase involves setting specific, realistic, achievable, feasible and measurable objectives taking into consideration the aims of metabolomics study on

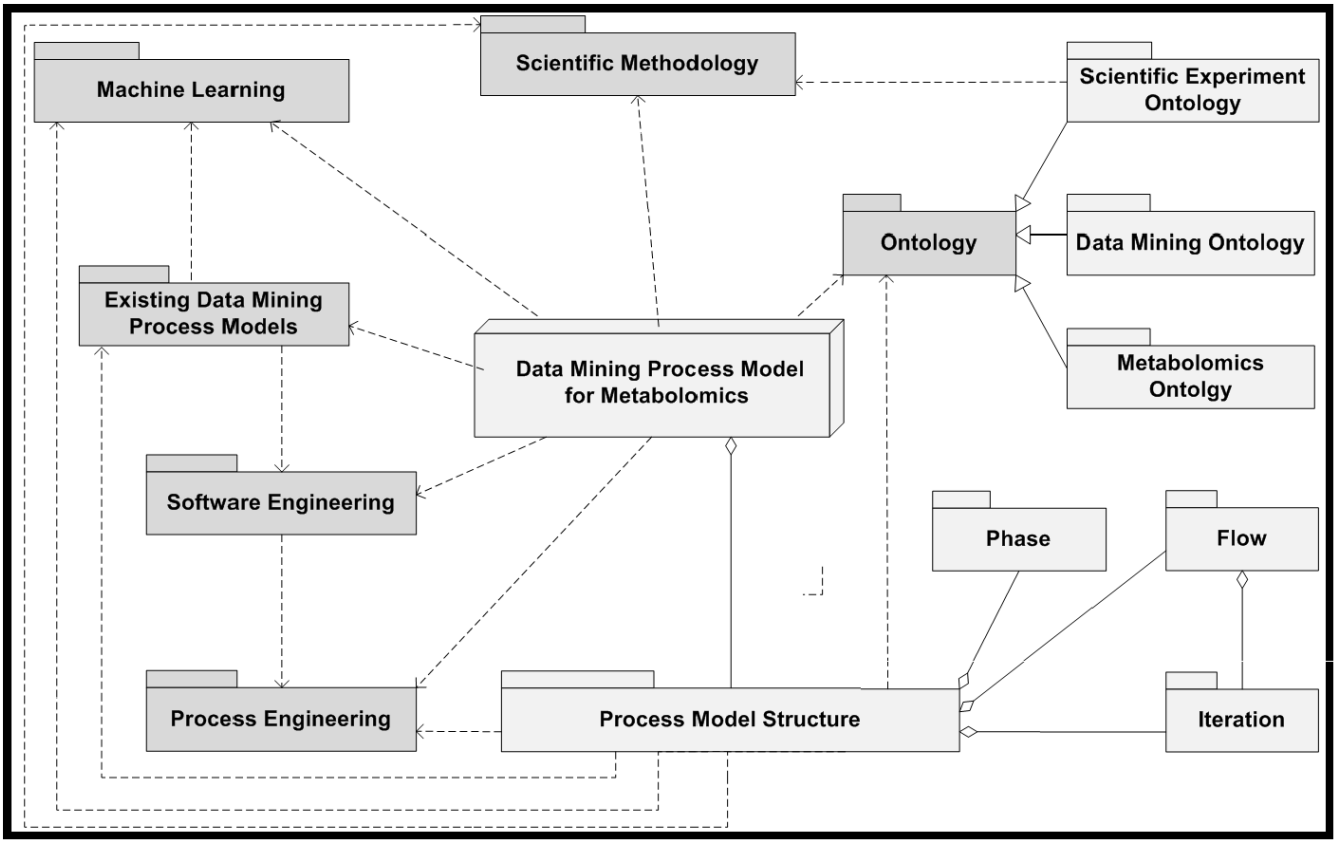

FIGURE 5. Process Model Design Foundations - A UML diagram illustrating the relationship between the process model and the foundations of its development. Dotted arrows represent dependency relationships between the proposed process model and its development inspirations and design foundations which are represented by the packages with darker colour. Triangular arrow heads represent a generalisation relationship that depict (is-a) relationships that relate to the ontologies of data mining, metabolomics and scientific experiments. Diamond arrow heads show aggregation relationships which depict (part-of) relationships between the process phases, flow, and iterations.

one hand, and the nature and potential of metabolomics data on the other. The aims of the study must be narrowed, and then expressed as data mining objectives, while still corresponding to the goals of the original metabolomics investigation. This phase was identified based on the scientific orientation requirement of the proposed process model and its design was inspired by the cycle of knowledge which requires formulating a testable and measurable hypothesis in order to obtain justifiable, traceable and reproducible results based on the principles and best practices of project management and software engineering which is used to drive software project planning, management and validation. Objective definition is common in most of the popular existing process models but its name, scope and context may differ from one to another.

Data Pre-Processing Phase aims at handling assay and other data acquisition related issues which may influence the nature and quality of acquired data. The activities of this phase are driven by the process defined objectives, as well as by the design of the metabolomics study. This phase was identified based on the preprocessing requirements of metabolomics data as described by the MSI and as discussed earlier in section III-A1.

Data exploration aims at understanding the nature and quality of the dataset, investigating its quality and prospecting its potential patterns, trends and correlations. The results of data exploration determines the exact data acclimatisation procedures to be applied and the data mining technique(s) to be selected and applied that suit the data and fulfills the process identified objectives. This phase was identified based on the requirement of metabolomics data analysis and its quality assurance and standardisation.

Technique selection aims to select the technique(s) that both achieves the process objectives and suits the metabolomics data. It involves matching the defined objectives to the goals, tasks, and data mining techniques, in order to select and justify the selection of the data mining technique that fulfills the defined process objectives and suits the targeted data. The selection is driven by the type of question that a metabolomics experiment intends to answer [44] and involves understanding data mining approaches, goals and tasks and the potential techniques and its application requirements and constraints e.g. software tools, packages, hardware, expertise, time, cost etc. The importance of this phase has been highlighted by both the data mining and metabolomics literature, while its design was based on a

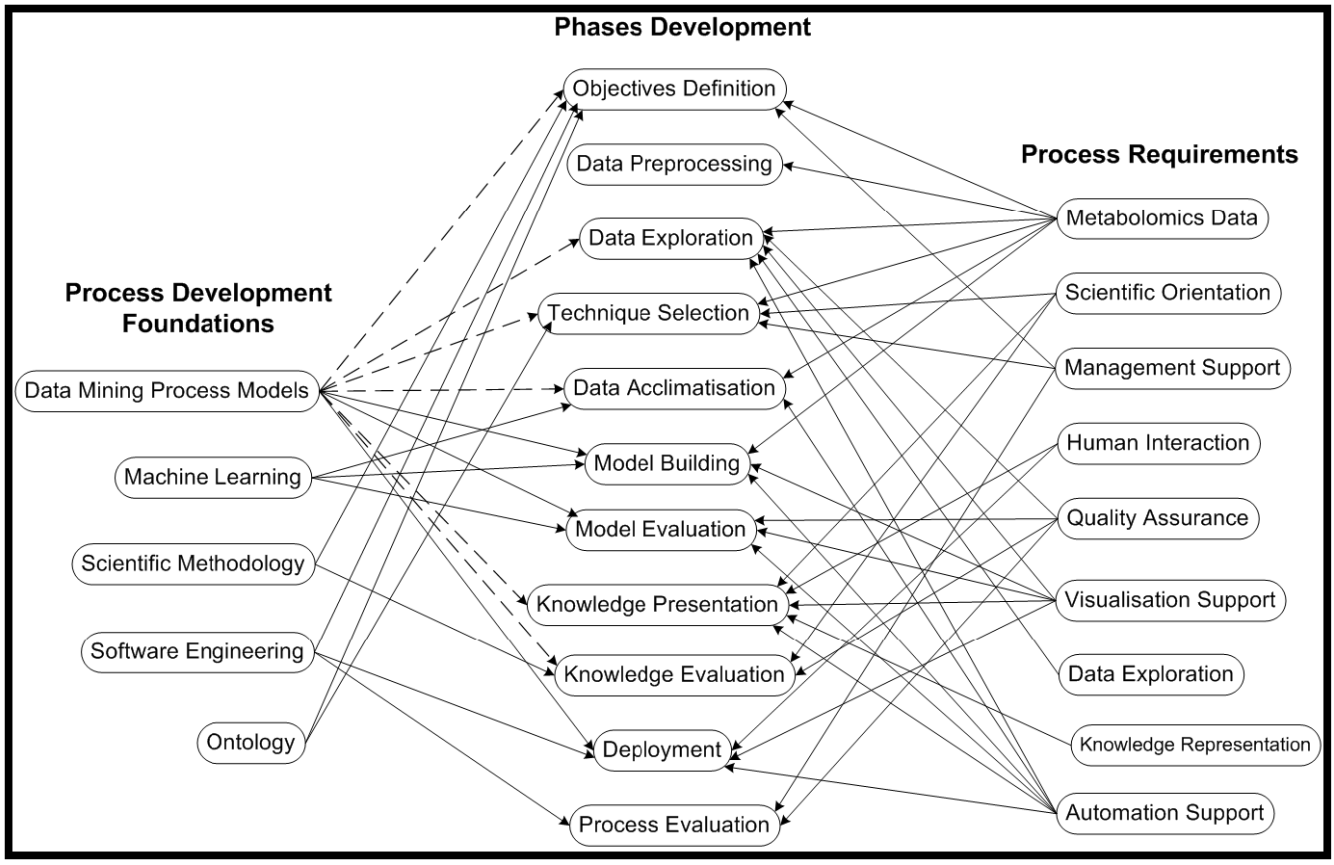

FIGURE 6. **Process Model Foundations to design Mapping- The mapping of the process requirements and development foundations to the identified process phases.**

strategy that published in [45]. Figure 6 provides an illustration of this strategy.

Data Acclimatisation phase aims to prepare the data to suit the requirements of the selected data mining technique. The identification of this phase was motivated by the pretreatment requirements of metabolomics data analysis and on the fact that data mining techniques which differ in their sensitivity and tolerances towards issues concern the nature and quality of the targeted data which may require proper handling such as size, high dimensionality, outliers, missing values, skewed distribution in addition to other data transformation, splitting, re-sampling, standardising procedures. However, the particular data acclimatisation procedures performed in this phase, must be driven by the needs of the selected data mining technique, and must be based on the defined process objectives, as well as the results of data exploration.

Model Building Phase aims to apply the selected technique to the targeted data in order to build the data mining model and to provide it with the proper training data. This phase is considered the backbone of the process model. It considers both technical and practical aspects and takes into consideration the modelling inputs requirements including data format, parameters and outputs; and the requirements of model evaluation, presentation and delivery as knowledge in later phases, as well as the process reporting standards. Model training ensures the validity of the data mining model. Training must be controlled by a stopping condition once the learning objectives are achieved [58].

Model Evaluation Phase aims to evaluate the model from a technical perspective. Model evaluation covers the assessment of model validity and performance, as well as the assessment of the model according to its measurability and success criteria defined in the objectives definition phase. This phase was identified based on the requirement of machine learning which requires assessing the correctness of the model and its ability to be generalised in order to describe, predict or classify unknown samples [83], [119]. However, the particular metric to be used in model evaluation depends on the technique(s) used for building the model as well as on the process objectives and their measurability criteria.

Knowledge Presentation Phase aims to express the data mining knowledge in an interpretable fashion using mechanisms such as interactive visualisation. This phase is important for knowledge understandability, interpretation, evaluation, utilisation, for answering **what-if** questions [52], [60] and also for tracing the knowledge presented to the underlying model inputs and parameters,for justifying the results obtained and for their reproducibility. It involves converting

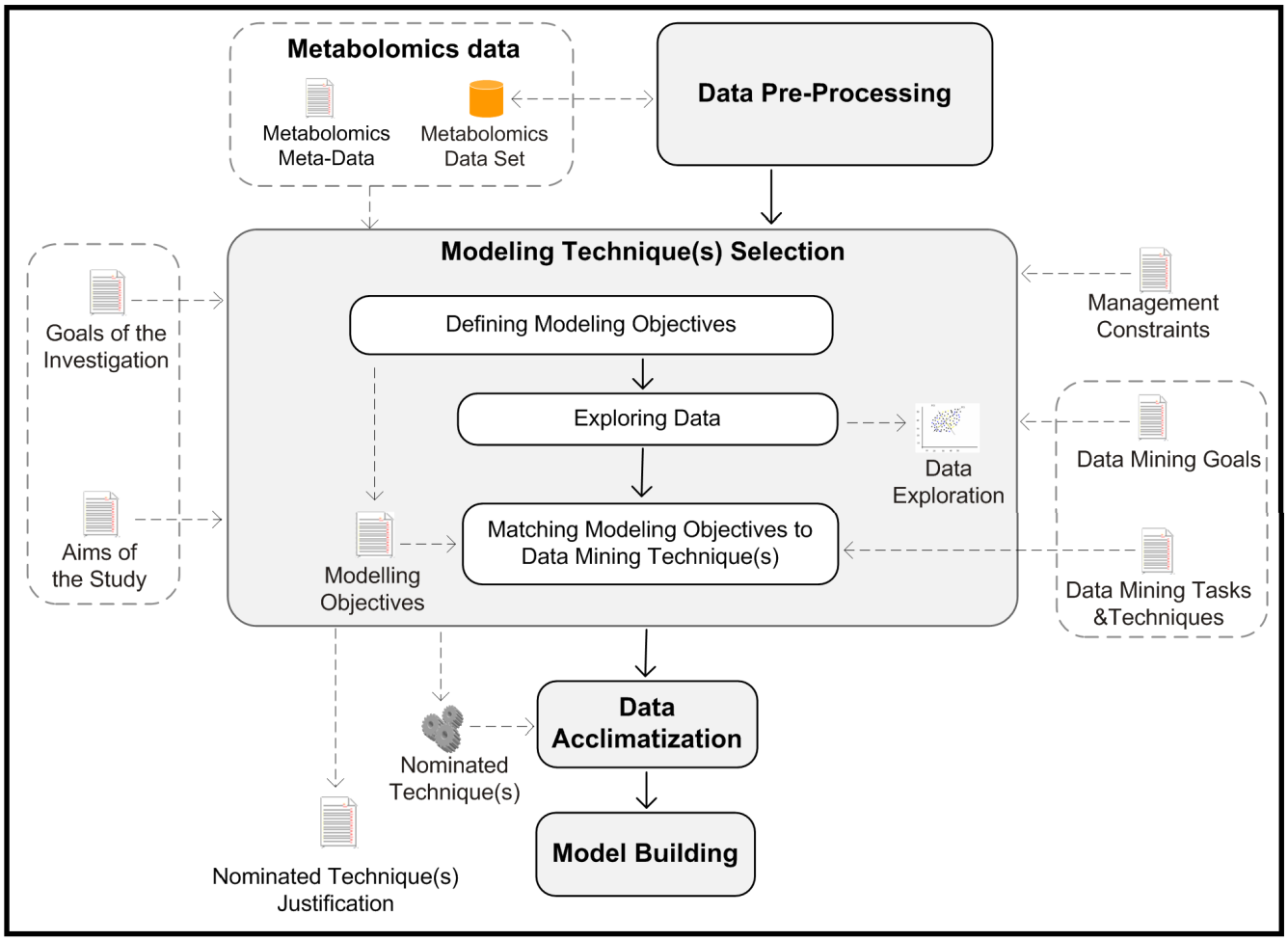

FIGURE 7. Illustrates the data mining technique(s) selection strategy that we proposed in [45].

information in the model from its machine learning or statistical context into a more meaningful depiction in the context of metabolomics that must be consistent with both the process defined objectives and the background knowledge.

Knowledge Evaluation Phase aims at evaluating the acquired knowledge in terms of its fulfilment of the process identified objectives and the aims of the metabolomics study in addition to its biological validity and usefulness [17]. It may involve evaluating the knowledge by a human expert based on the existing body of knowledge or with other results generated by other metabolomics investigations. This phase was identified based on the scientific nature of metabolomics investigations and also on the results validation requirements of metabolomics data mining process. The activities in this case were based on the principles of scientific methodology, as well as on verification and validation concepts in software engineering and machine learning. The outcome of this phase might either confirm with or contradict the background knowledge or findings of other researches and might lead to discovering a new phenomenon, pose a new question or generate a new hypothesis [21]. Although the activities of this phase are semantic, it must avoid bias and encourage consistent decisions through automation which is quite challenging under the current technology.

Deployment Phase aims to facilitate knowledge accumulation, utilisation and further analysis and interpretation in the context of new discovered knowledge [22], [58]. The deployment phase was based on the requirements of the metabolomics data mining results interpretation and utilisation and the process reporting standards [63]. This phase was inspired by the results deployment features available in some of the existing data mining process models, e.g. The CRISP-DM, EBPM and by knowledge utilisation features available in other process models, e.g. the KDD. Several options are available for deploying data mining results including software applications, databases, or reporting. The selection of the deployment mechanism must consider the type of knowledge to be deployed, its underlying modelling technique, its potential usefulness [120] and its satisfaction of the reporting standards e.g. MeMo [48].

Process Evaluation Phase aims to ensure the quality of the results and the correctness, completeness and validity of the

procedures applied, decisions involved and the deliverables generated throughout the process either on the micro level of process phases, or on the macro level of the process execution. Process evaluation ensures the process execution compliance with the flow of the process phases layout and also, with the internal tasks within phases and the consideration of the practical requirements of the process model including management, human interaction, quality assurance and reporting standards. This phase was identified based on both the requirements of metabolomics and data mining validation and quality assurance requirements. It includes features inspired by both data mining and software engineering process models e.g. cross-deliveries validation [101], [102].

2) LAYOUT STRUCTURE AND DESIGN

The design of the proposed process model builds on the good features of the existing data mining process models and addresses their gaps, weaknesses, and shortcomings. The process model also provides improvements in terms of its layout and design structure, which are based on the solutions inspired process engineering, scientific methodology, software engineering and machine learning. However, the process model development was still driven by the requirement of metabolomics data mining. Figure 8 illustrates the relationship between the process model design and its development requirements and foundations.

A prototype of the process model was developed and improved through several iterations, validations, and verification which focused on satisfying the requirements of metabolomics data mining. The process layout structure considered the principles of process engineering and software engineering regarding process models design [97] and best practices including abstraction, notation, content, modularity, instantiation, cohesion and understandability [96].

The process model was laid out in a well-structured and well-organised fashion that increases the coherence of the tasks within the process phases in order to reduce the complexity of the process model and to enhance the modularity of its phases, which is an important principle in process engineering.

The flow of the process phases was organised based on the workflow of metabolomics data analysis and the generic data mining process phases. It defines the process flow and iteration, as well as the relationships between its phases, either in terms of their flow to their successors or in terms of their feedback to their predecessors. This enhances the process applicability and improves the validity of its results and the flexibility in its execution.

The layout structure of the process model was improved based on ideas inspired by a several aspects of software engineering methodology which include: (1) v-Model cross-deliveries validation where the deliveries of the earlier phases are checked by the deliveries of the relevant later phases; (2) the validation of data, deliveries, and the phases internal activities;(3) the process evaluation, knowledge evaluation and model evaluation which is also pivotal in machine learning;(4) the iteration and feedback support; (5) the principles of coupling and cohesion; (6) the justification and traceability support which is also emphasised in the principles of the scientific methodology; (7) consideration of the practical aspects including management, quality, standards and human interaction.

D. SOFTWARE ENVIRONMENT IMPLEMENTATION

The idea of providing a software realisation of the data mining process was inspired by the Rational Unified Process in software engineering [65]. The process model was realised in a software environment called MeKDDaM-SAGA which was created in order to implement and realise the proposed process model and automate its features as well as to guide the flow of its phases and aid its execution. The software was constructed as a visual GUI environment based on the principles and best practices of object-oriented Software Engineering which covered the analysis of the software requirements, its design models and architecture in addition to its construction and testing. The environment was designed to enable the execution of the process model either externally using independent software tools and then recording the executed process activities and importing their outcomes to the environment, or internally using a number of embedded tools and facilities e.g. artificial Neural nets, decision trees, random forest, hierarchical clustering, etc. This covers data exploration and acclimatisation in addition to model building, model evaluation, and knowledge presentation phases.

The process model realisation software helps guiding the execution of the process normal flow, feedback, and iteration and the execution of tasks within the process phases including prerequisites, objectives, planning, performing, validation and reporting. The implementation software realises the process model support to the practical considerations including management, human interaction, quality assurance and standardisation. It also helps the realisation of the process scientific orientation through providing a number of various traceability, justification and validation mechanisms. The requirements of the software environment have been identified based on software engineering methodology and in the light of the proposed process model requirements, foundations, development and description as discussed in the previous sections.

MeKDDaM-SAGA provides support for metabolomics data requirements, scientific orientation and for practical considerations and other desired features. It also offers several facilities that enable the execution of the process model either externally or internally using various embedded facilities which enable building and evaluating data mining models using a number of popular machine learning techniques that include: (1) Random Forest; (2) Support Vector Machines (SVM); (3) Hierarchical Clustering Analysis (HCA); (4) Association Rules; (5) Bayes Nets; (6) Principal Component Analysis (PCA); (7) Multi-Layer Perceptron Neural Networks (MLP); (8) Decision Trees, in addition to the visualisation of last two. The software implements an internal

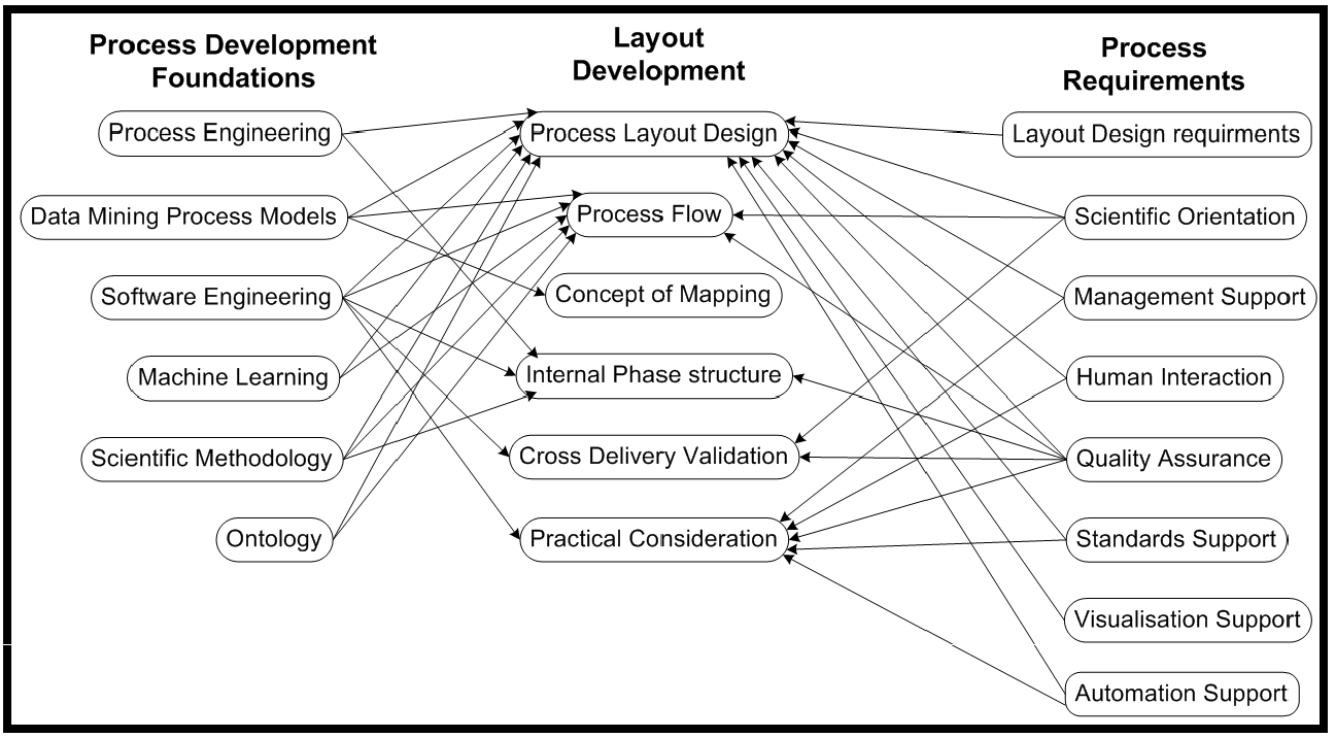

FIGURE 8. **Process Model Foundations to design Mapping- The mapping of the process requirements and development foundations to the process model design.**

algorithm to version control the process execution feedback and iterations and to enable the undo and redo of every aspect concerning the execution.The software persists its data using XML in order to grantee the maximum level of portability and a GUI to ensure its easy of use without prior training. The software was implemented in Java and constructed based on the principles, methodologies and best practices of object-oriented Software Engineering. The software was designed to realise 27 use-cases and implemented using 27 packages which are shown in Figure 9, while Figure 10 shows a UML class diagram that was used to realise and automate the process phases and all internal tasks. More details regarding the development of the software implementation environment are described in an ArXiv preprint [15], while MeKDDaM-SAGA software is freely available as an open-source software at GitHub (https://github.com/banimustafa/MekDDaM-SAGA) [16].

E. PROCESS MODEL DEMONSTRATION APPLICATIONS

MeKDDaM was applied to four metabolomics applications to demonstrate its applicability and to evaluate its execution in the context of its development requirements. The applications covered plant genetics (***Arabidopsis thaliana***), animal nutrition (***Dairy Cows Diet***), and human disease (***Kidney Disease***). The applications provided coverage of the three major metabolic approaches including metabolite profiling, targeted analysis, and fingerprinting and their datasets that were captured in assays that involves samples from plant, animal and human origin using the three major groups of techniques: chromatographic separation and MS-Based technique (LC-MS), optical technique (FT-IR) and Nuclear Magnetic Resonance (NMR). The applications provided coverage of both data-driven and hypothesis-driven data mining approaches and demonstrated the process' ability to fulfill a variety of objectives derived from the three major data mining goals: prediction, description and verification in addition to carrying out a range of data mining tasks including classification, segmentation, hypothesis testing, correlation analysis, dimensionality reduction, feature extraction and hypothesis testing. However, some of these tasks were applied only in order to perform exploratory data analysis as part of data prospecting such as correlation analysis and dimensionality reduction while others were performed for the purpose of **model building** using a number of example data mining techniques i.e Artificial Neural Networks (ANN) [121], Self Organising Maps (SOM) [122], Support Vector Machines(SVM) [123], [124], Principal Component Analysis (PCA) [125], Hierarchical Clustering Analysis (HCA) [126], Partial Least Square Discriminant Analysis (PLS-DA) [127], decision trees [128] and random forests [129]. Figure 11 summarises the applications discussed in this work and illustrates their coverage of metabolomics instruments and approaches as well as their coverage of data mining approaches, goals and tasks. More details regarding the process model demonstration applications is available in [14].

F. PROCESS MODEL EVALUATION

The process model was evaluated in light of the identified requirements of metabolomics data analysis and data mining

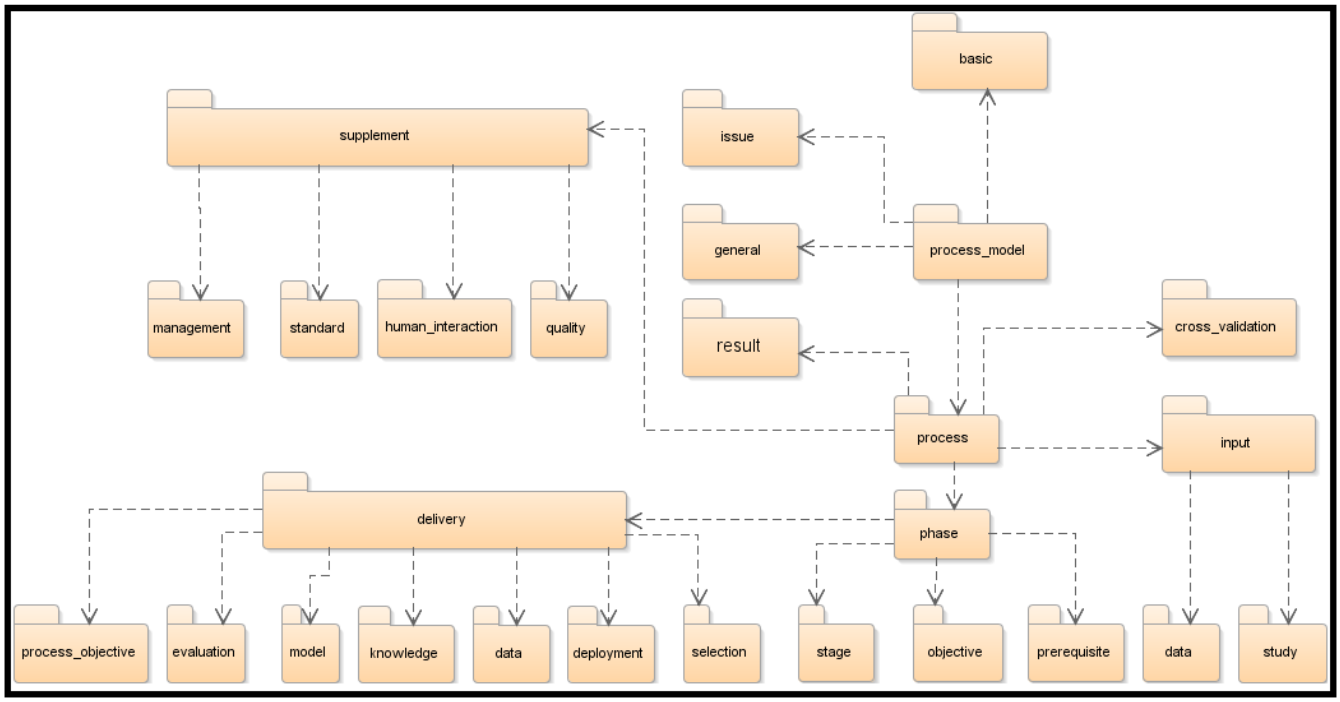

FIGURE 9. **Software Packages:** A UML diagram showing the packages in the software environment used for implementing the process.

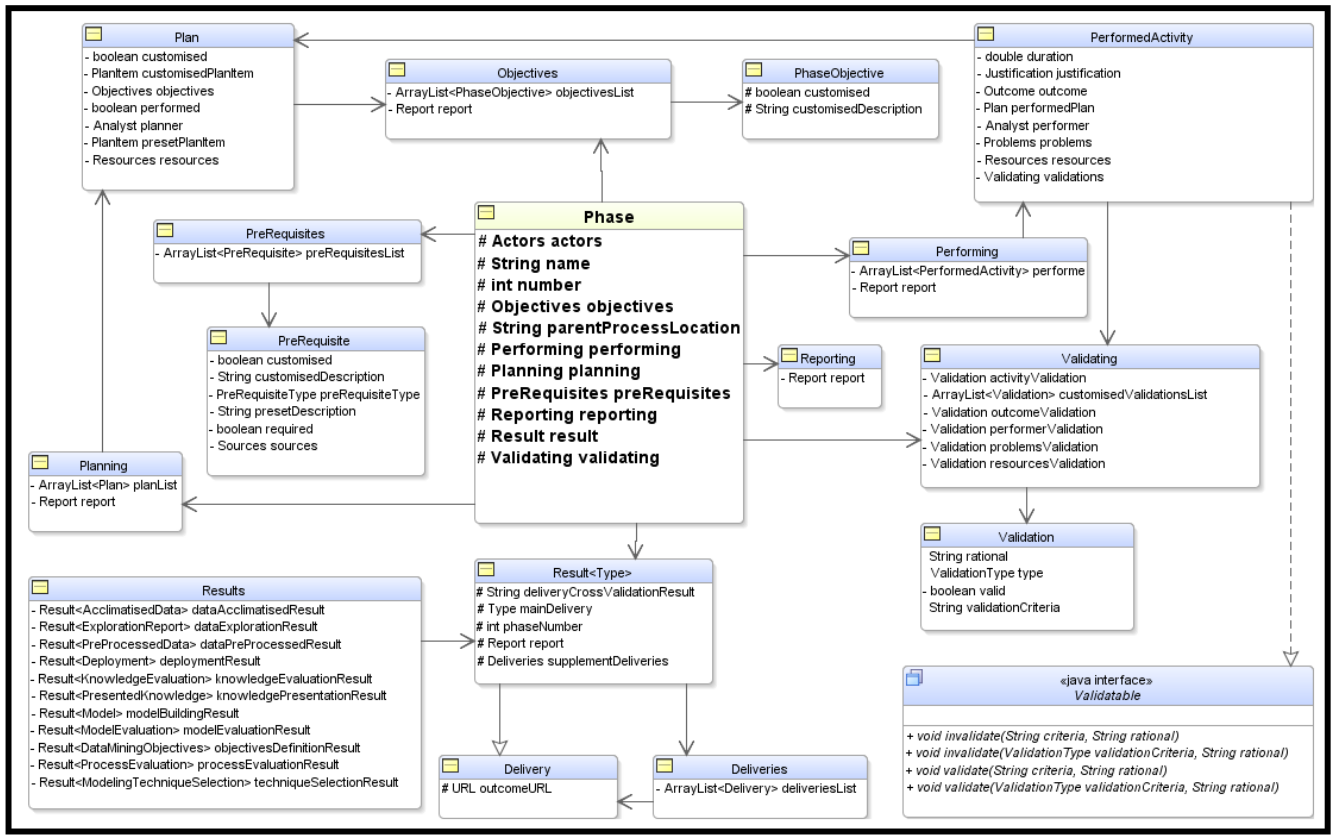

FIGURE 10. **Phase Class Diagram:** A UML diagram representing the process phases and all relevant classes as implemented in Java. Sharp head arrows represent a dependency/usage relationship between classes, while rectangle head arrows represent an inheritance relationship between classes which depict is-a relations. Dotted arrows with a rectangle head are used to represent realisation/implementation relationships between classes and interfaces.

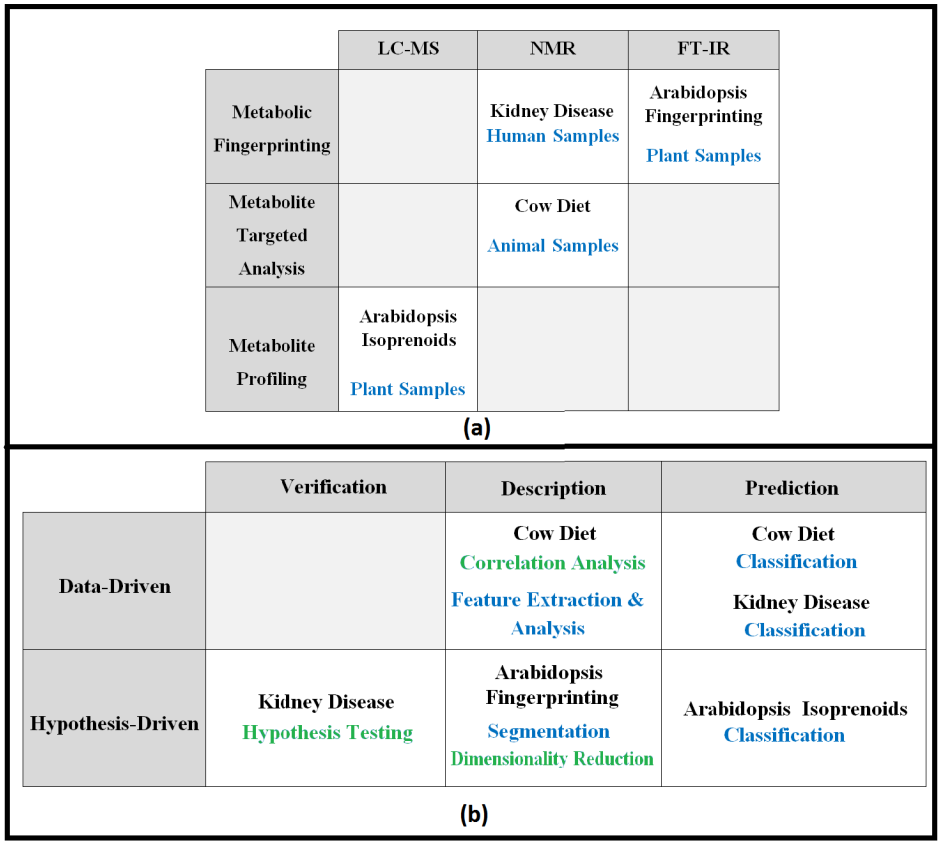

FIGURE 11. **The Process Demonstration Applications- (a) The applications' coverage of the metabolic approaches (rows) and the popular applied Data Mining Techniques (columns) (b) The applications' coverage of data mining approaches (rows), goals (columns) and tasks (cells). The tasks in blue were used for model building, while the ones in green were used as part of the exploratory analysis which was carried during data prospecting.**

process in addition to the foundations and inspiration of its enhancement and improvement, and based on the outcomes of the process demonstration using the implementation software environment and the four real-life applications.

The scientific orientation was demonstrated through applying the traceability and justification mechanisms described by the process and realised in the software environment which contributed towards the consistency of the analysis results. The process execution also confirmed the validity of the process flow and structure and the validity of the prescribed tasks within the process phases. The process phases and their internal tasks were found to be both cohesive and understandable, thanks to the process concept of mapping, where the phases generic tasks were customised to suit the needs of the application as well as the process practical supplements and traceability. The process concept of mapping through its implemented customisation mechanisms was vital for improving the process efficiency and agility, which saved a considerable time during the process execution. The software allowed loading of the process phases objectives, prerequisites, and planning based on the process model description and provided facilities for customising these tasks for the specific needs of each application. The software also supported the reuse of the process customisation, which can be saved and then reloaded by similar applications. The software customisation facilities support storing, exporting, and importing the customisation of the process phases as well as for the customisation of the process practical supplements and traceability.

The process iteration was found useful for re-executing the process in order to answer a different question as demonstrated in the cow diet application. The rollback mechanism, described by the process model and implemented by its realisation software was also found useful. The process model iteration, feedback, and rollback mechanisms were also practically useful. They helped organise the forking of the process flow, while maintaining the validity and consistency of the execution of its phases and deliveries. Phase feedback and iteration were particularly useful for building alternative models as well as for updating the execution of the process phases, enhancing their performance or improving the quality of their deliveries.

The process execution demonstrated support of management and planning, which was illustrated through the various

planning and management mechanisms realised by the software environment either on the level of the process, or on the level of internal tasks within its phases. The process execution demonstrated the realisation of the process resources management and allocation which was considered in almost every aspect of the process. Human interaction was demonstrated using the mechanisms provided by the software environment, such as assigning the performer of the process activities as well as assigning humans as a traceable source using the traceability and justification mechanisms. The execution of the process demonstrated the various quality assurance mechanisms provided by the process and its implementation software in order to satisfy the requirements defined.

The quality of the data is investigated in the data exploration phase and handled later in data acclimatisation as demonstrated in the kidney disease application. In addition, the validity of the model is evaluated in the model evaluation phase as demonstrated in all the four applications. The process application also demonstrated the benefits of considering metabolomics and data mining standards in the development of the process model. The deliveries of the process phases were designed to comply with the MSI reporting standards. This was demonstrated by the application through the performance of both data preprocessing and pretreatment procedures, which have been considered in the designing of the data preprocessing and data acclimatisation phases as well as in their realisation by the software environment. However, the process execution support for reporting standards e.g. PMML depended on the Data Management Group (DMG) support of the particular technique used for model building as DMG provides support only for some of the data mining techniques.

The applications demonstrate the process support for visualisation as a preferred tool in knowledge presentation. Visualisation was found particularly useful for data understanding, investigation and prospecting in data exploration, technique selection and data acclimatisation phases. The applications also demonstrated the process model's satisfaction of the knowledge presentation requirements as well as its realisation by the process software environment. The importance of knowledge presentation. Furthermore, the applications demonstrated that the software realisation satisfies the automation requirement and that the demonstrated applications performed all the process model execution scenarios discussed earlier.

The process evaluation results concluded that the process outperformed all the existing data mining process models and confirmed its satisfaction of the requirements of metabolomics data mining. It also confirmed the process support to a number of practical aspects in respect to manageability, human interaction, quality assurance, and both metabolomics and data mining standards. The results also confirmed the process support of several desirable features including visualisation, data exploration, knowledge presentation, and automation and highlighted a number of unique features that were inspired by: process engineering, software engineering, machine learning and fundamentals of scientific methodology. Furthermore, the proposed process model offered major contributions toward the improvement and enhancement of the data mining process in general and more particularly in scientific applications.

IV. RESULTS

The proposed process model (MeKDDaM) describes the process layout structure, phases, actors, inputs, and deliveries. It defines the normal process flow, feedback and iterations, as well as showing the process practical considerations namely: quality, management, standardisation and human interaction. The process model consists of eleven phases, which are organised in a v-shape, in order to be executed sequentially and iteratively. All process phases have the same internal tasks, which cover their prerequisites, objectives, activities planning, performing and validation, as well as reporting. The process defines the normal flow of its phases, as well as its execution iteration and the feedback between its phases in addition to the participants in each phase execution. The inputs and outputs of each phase are also defined in the process model both on the macro level the process execution as whole and also on the micro level of internal tasks running within phases. The inputs for the process include the aims of metabolomics study, the targeted dataset and its associated meta-data, while the input phases are the outputs of their prior phases. The deliveries of each phase are reported as inputs for its successor, or as part of the intermediate or the final results of the process. The output of the process execution on the macro level is the deployable knowledge and its associated model(s), while the outputs of each phase on the micro level cover both its generated outcomes and the reporting of its internal tasks running. The MeKDDaM process model is illustrated graphically in Figure 12.

A. INPUTS OF THE PROCESS MODEL

Process inputs include metabolomics data, to which the process model is applied, as well as the aims of the study in the case of hypothesis-driven data mining. The metabolomics data must be exported into a persistence mechanism that allows accessing, storing and handling metabolomics data. This could be a database, files or other data repositories. When sampling the targeted data or using a subset of the metabolomics data, the subset must be validated to ensure that it represents the targeted data, and that it is sufficient and appropriate to be analysed using data mining techniques in general, in order to fulfil the defined process objectives

B. LAYOUT OF THE PROCESS MODEL

The process model describes the process layout structure, phases, actors, inputs, and deliveries. It also defines the normal process flow, feedback and iterations, as well as showing the process supplements: quality, management, standardisation, and human interaction. The process model consists of eleven phases, which are organised in a v-shape, in order to be executed sequentially in an iterative fashion. The process

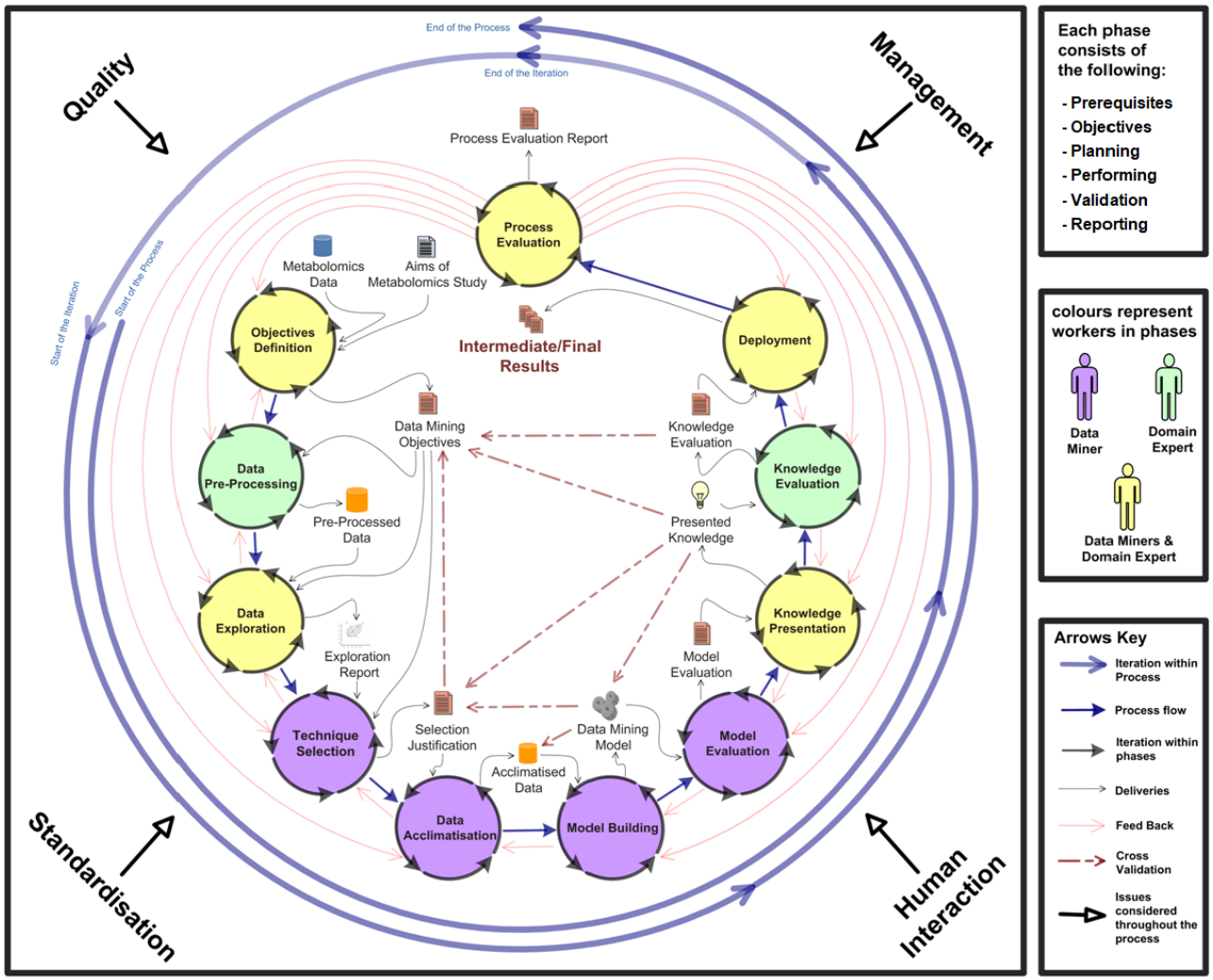

FIGURE 12. MeKDDaM Process Model Graphical Representation- The black arrows show the flow of process phases (counter clockwise), while the orange arrows show feedback between its phases(clockwise). The dashed arrows show the cross-validation relationship between deliverables. The black circular arrows around the process phases show the iterative nature of their internal tasks running, while the blue spiral arrows around the process shows the iterative nature of the overall process model execution. The process phases are colour-coded to show the role of participants in the execution of each phase, while the practical supplements on the corners illustrates their embedded consideration throughout the process execution as indicated by the triangle headed arrows.

phases have the same task templates, which covers their prerequisites, objectives, activities planning, performing and validation, as well as reporting. It defines the normal flow of the process model, as well as its iteration and the feedback between its phases. The inputs and outputs of each phase are also defined in the process model, as well as the participants in its execution. The deliveries of each phase are reported as inputs for its successor, or as part of the intermediate or the final results of the process. The design of the process model is illustrated in Figure 12.

The process execution is performed by running its phases, taking into consideration their inputs and deliveries. Each of the process phases defines an iterative sequence of tasks which must be performed when running the phase, either as part of the process normal flow, or in phases feedback and process iteration. Process iteration is defined as the repetitive execution of all process phases from the beginning to the end, maintaining the order and flow of phases, and taking into consideration their defined inputs and deliveries. Feedback is a micro scale iteration which involves two or more of the process phases. All the phases included in the feedback ring must be revisited in order to maintain the process flow, and to ensure the consistency of their deliveries.

The process model also integrates its support for data mining practical aspects by embedding them in its layout design and in the description of its phases. Human interaction is considered in all phases, either on the level of process phases execution or on the level of the tasks within its phases. The process uses a colour coding system to represent the participant in each phase, i.e. data miner, domain expert or both. The

process structure and description also provides details regarding who is doing what, where, and how. Management issues are considered on the level of the process, as well as on the level of its phases. The objectives of the process are defined in the first phase in terms of their measurability, success criteria, and feasibility. These are used later for technique selection, and for the evaluation of the discovered knowledge and process execution. Management aspects are in the structuring of the tasks within the process phases, where the objectives of the phase are set, and its planned activities are executed, validated and reported. Quality assurance is embedded in the design of the process model and also considered in its execution. The process model validates the quality of the data, as well as the model and its presented knowledge. The execution of the process model is also validated, as well as activities within its phases. The process model also provides an extra level of validation on the level of its deliveries. Metabolomics and data mining standards were both used for designing the process layout, as well as their ontologies. The process also encourages use of the reporting standards in the delivery of its phase, e.g. PMML.

Additionally, the concept of mapping in CRISP-DM has been utilised in the development of the data mining process model for metabolomics. Using this concept, the process model can be broken down into a group of stages. Each consists of a set of the generic tasks of the process model, and each is customisable and can be specialised to be applied to a particular application. The process model defines a mechanism for evaluating process deliveries in terms of backward report-level consistency check. Cross-delivery validation was inspired by software engineering v-model. It aims to provide a high level quality assurance for the process on the level of its generated deliveries. The results of cross-delivery validation can cause feedback to resolve the inconsistencies, as well as process iteration.

C. PHASE INTERNAL TASKS

The process phases are designed to be performed by executing a number of iterative tasks that include:

- **Phase Prerequisites:** encompasses the inputs delivered by the previous phases, process inputs and the background knowledge and information recorded and considered for phase customisation, implementation, and running. Phase prerequisites must be valid, specific, relevant and sufficient to run the phase without the need for additional resources. This enables the justifiability and traceability of the phase results, as well as the reproducibility of its validated deliveries.
- **Phase Objectives:** defines the operational objectives the phase and its expected deliverables and their desired attributes and characteristics, which are defined in a fashion that is analogous to the concept of functional and non-functional requirements in software engineering.
- **Phase Planning:** concerns the mapping of the phase objectives to a set of practical actions, designed as a sequence of activities that aims to fulfil the phase objectives. The planned activities take the phase prerequisites as inputs in order to generate the phase deliverables. Phase planning must comply with both data mining and metabolomics procedural standards, and it must also be also in line with project management principles, and human interaction best practices.
- **Phase Performing:** involves carrying out the phase planned activities and all its related decisions which are justified and recorded along their relevant evidences. The problems, gaps and limitations encountered during the running of the phase activities are also recorded to be reported and considered in phase validation and in future phase and process iterations as well as in cross-deliveries validation and process evaluation.
- **Phase Validation:** aims to ensure the quality of the phase activities performance and the quality of the data involved and its compliance with the standards adopted. It validates the performed activities according to the phase plan and according to their fulfilment of the phase objectives.
- **Phase Reporting:** concerns the generation of the phase outcomes, processed data, and other deliverables, as well as the reporting of the phase internal tasks running. Phase outcomes and deliverables must conform to the applicable standards in both metabolomics and data mining where possible e.g. PMML, XML, ArMet, MeMo

D. PHASES OF THE PROCESS MODEL

MeKDDam consists of eleven phases. The scope, context and rational of each phase was discussed earlier in Section III-C1. Here, we provide details regarding the process phases and their internal tasks.

1) OBJECTIVES DEFINITION

This phase provides a mechanism for defining the process modelling objectives based on data mining approaches, goals, and tasks as derived from the aims of the metabolomics study and their relationship with the goals of its original research investigation.

a: PHASE PREREQUISITES

- The metabolomics dataset and its associated meta-data if available.
- The aims of the metabolomics study and its investigation goals, hypotheses, and assumptions.
- Background knowledge/information regarding data mining including its approaches, goals, tasks, and techniques.
- Background information regarding project management and planning.
- The process management constraints, including: time, cost, and expertise as well as the availability of the software and hardware infrastructure.
- The process quality assurance policy.
- The process standards.

b: PHASE OBJECTIVES

- Defining the data mining modelling objectives for metabolomics.
- The objectives should be expressed, either in a hypothesis-driven or data-driven fashion. Hypothesis-driven objectives are derived from the aims of the metabolomics study, while data-driven objectives are derived from the general goals of data mining and tasks.
- The objectives must be realistic and achievable using data mining techniques in general and using the targeted metabolomics data, which must be sufficient, adequate, and relevant to the aims of the metabolomics study either to test an implied hypothesis, or to discover new knowledge.
- The objectives must be feasible. It must take into consideration the process management constraints, including: time, cost, and expertise as well as the availability of the software and hardware infrastructures.
- The fulfilment of the objectives must be measurable and the success of their achievability must be testable.
- The objectives must be flexible and adjustable in case of altering the process flow as a result of a feedback or iteration.

c: PHASE PLANNING

1) Review the aims of the study and their relationship with the investigation goals, hypotheses, and assumptions.
2) Verify the general understanding of data mining goals, and tasks as discussed earlier.
3) Decide the type of the intended objectives, whether they will be hypothesis-driven, based on the aims of the metabolomics study, or data-driven, based on the general understanding of data mining approaches.
4) Derive the modelling objectives depending on the type of data mining intended:
    a) Hypothesis-driven objectives should be derived from and consistent with the goals of the original investigation, the research hypothesis and assumption, and the aims the metabolomics study:
        i) Analyse the relationship between the data mining approaches, goals, tasks, and techniques on one hand, and the aims of the metabolomics study on the other hand.
        ii) Translate the aims of the metabolomics study into definable data mining modelling objectives, based on the general understanding of data mining goals and tasks.
    b) Data-driven objectives should be derived from the general goals of data mining goals and its subsequent tasks.
        i) Analyse the potential of the targeted dataset and its associated meta-data (if available) for fulfilling possible data mining goals and tasks.
        ii) Translate the data mining general goals and tasks into narrow and more specific data mining objectives.
5) Assess the derived objectives achievability using the available data mining techniques and the targeted metabolomics data in terms of its sufficiency, adequacy, and relevance.
6) Assess the derived objectives feasibility in the light of the process management and technical constraints.
7) Depending on the results of the assessments in steps 5-6, decide whether to go to step 8, or to consider alternative objectives by going back to step 4.
8) Define the data mining objectives of the process and their success criteria.

d: PHASE PERFORMING

1) Perform and record the planned data exploration activities.
2) Justify and record the chosen phase options and alternatives.
3) Document the problems, gaps, and limitations that are encountered when carrying out the planned phase activities.

e: PHASE VALIDATION

1) Validate the defined objectives in terms of their correctness, completeness, and consistency with the aims of the study, goals of the investigation, research question, hypothesis and assumptions on one hand, and with the goals of data mining and tasks on the other hand.
2) Validate defined objectives in terms of: achievability, measurability, testability, and their technical implementation feasibility in terms of time, cost and the availability of expertise, software, and hardware infrastructures.
3) Validate that the data is relevant, sufficient, and adequate to fulfil the defined objectives.
4) Validate the achievement of phase objectives.
5) Validate that all the activities in the phase have been carried out, recorded and justified according to the plan.
6) Validate that all the problems, gaps and limitations encountered in the phase activities has been recorded according to the plan.

f: PHASE REPORTING

1) Define the applicable reporting standards based on the available standards in data mining and metabolomics.
2) Based on the defined standards, report: the defined data mining modelling objectives; success criteria; deriving mechanism; assessment and its associated justification information, including the type of the modelling objectives (hypothesis-driven fashion or data-driven); the traced origin of the modelling objectives (a reference to the aim of the study and goal of the investigation it hopes to achieve or to the general goal of the data

mining it was inspired by); the objectives measurements; and success criteria.
3) Report the human interactivity as this phase is shared by both data miner and domain expert. This is done by reporting the performer of each of the phase performed activities.
4) Report the procedures carried out and as well as their involved justification, in the case of deciding alternatives and options.
5) Report the phase validation outcomes.
6) Report the problems, gaps, and limitations, which might be encountered during the performance of the phase activities.

2) DATA PRE-PROCESSING

Data preprocessing is an optional phase which aims to clean the raw data as acquired by the assay instruments. The scale of this phase, and its particular applied procedures, depend on preprocessing procedures performed during data acquisition, as well as on the defined process objectives, and the design of the metabolomics study.

a: PHASE PREREQUISITES

- The metabolomics dataset and its associated meta-data if available.
- The defined objectives of the process.
- The aims of the metabolomics study.
- Background information regarding the design of the metabolomics study and assays.
- Background knowledge regarding metabolomics in general, metabolic approaches, design of metabolomics study, data acquisition techniques and data preprocessing procedures in general.
- The design of the metabolomics study and the assay which was used for acquiring the data. This must include details regarding the bio-sample and sampling protocols, as well as sample preparation procedures and data acquisition including any preprocessing procedures already performed on the level of the assay.
- The process quality assurance policy.
- The process standards.

b: PHASE OBJECTIVES

- Pre-processing the metabolomics data.
- The data preprocessing must be carried out in the light of the modelling objectives and should be consistent with the design and the aims of the metabolomics study and their relationships with the goals of the investigation, hypothesis and assumption.
- The data preprocessing procedures must be based on the nature and the requirements of the metabolomics data as generated by the data acquisition technique and must take into account, the metabolic approaches, the design of the metabolomics study and its subsequent assays which were used for acquiring the targeted data.
- The requirements of the data preprocessing procedures must be considered including: time, cost, expertise and software and hardware infrastructures.
- The data preprocessing procedures must be comprehensive, adequate, and compatible with the requirements of the data acquisition assay, and must take into consideration the preprocessing procedures, which were already performed by the data acquisition.
- The data preprocessing procedures must be adjustable to reflect any change in the case of feedback from later phases or in case of process iteration.

c: PHASE PLANNING

1) Review the defined process objectives and their relationship with the aims of the study.
2) Identify the required data preprocessing procedures, based on the design of the metabolomics study.
3) Identify the requirements of the identified preprocessing procedures.
4) Perform the data preprocessing procedures and record their purpose and description.

d: PHASE PERFORMING

1) Perform and record the planned data exploration activities.
2) Justify and record the chosen phase options and alternatives.
3) Document the problems, gaps, and limitations that are encountered when carrying out the planned phase activities.

e: PHASE VALIDATION

1) Validate that the preprocessing activities carried out have fulfilled the phase set objectives, in terms of their comprehensibility, adequacy, adjustment flexibility and compatibility with the nature and requirements of the preprocessed data as they have been acquired by the assay.
2) Validate that all the preprocessing phase activities have been carried out, recorded, and justified according to the plan.
3) Validate that all the problems, gaps and limitations encountered in the phase activities have been recorded according to the plan.

f: PHASE REPORTING

1) Define the applicable reporting standards, based on the available standards in data mining and metabolomics.
2) Report the preprocessed data.
3) Report the procedures carried out and as well as their involved justification, in the case of deciding alternatives and options.
4) Report the phase validation outcomes.
5) Report the human interactivity as this phase is shared by both data miner and domain expert. This is done by reporting the performer of each of the phase activities.

6) Report the problems, gaps, and limitations, which might be encountered during the performance of the phase activities.

3) DATA EXPLORATION

This phase involves performing a set of activities which aims to get insight into the data and contribute towards the selection of the data mining modelling technique. The activities of this phase include: data investigation, understanding and prospecting. The output of this phase takes the form of a report, which contains details regarding the activities that are performed in the phase and their outcomes.

a: PHASE PREREQUISITES

- The preprocessed metabolomics dataset and its associated meta-data if available.
- The defined objectives of the process.
- The aims of the metabolomics study.
- A background knowledge/information regarding the assay including its data acquisition technique, metabolic approaches, and study design.
- Background knowledge/information regarding the basic statistical measures and data visualisation.
- The process quality assurance policy.

b: PHASE OBJECTIVES

- Examining the nature of data, e.g. data types, structure, size, files and data format.
- Understanding the data by describing the meanings of data attributes and the range of their values, as well as the relationship between their variables.
- Investigating the quality of metabolomics data, depending, either on the specific quality assurance standards provided as prerequisites to the phase, or on the general data quality standards defined in data mining, regarding data distribution, and the existence of missing values and outliers.
- Prospecting the potential of the data to address the objectives and interesting distributions, trends and relationships. This can be carried out using the available visualisation tools, basic statistical measurement, and other techniques in order to provide a comprehensive overview of data from different perspectives.
- Verifying the sufficiency, adequacy and relevancy of metabolomics data to fulfil the defined modelling by investigating their nature and explaining the meanings of their attributes and values.
- All data exploration procedures must be comprehensive and thorough, and must cover the entire data, and encompass all their related aspects.
- All data exploration must be carried out in the light of the defined process objectives, and in the case of hypothesis driven data mining,must also be in line with the aims of the metabolomics study and with the goals, hypothesis and assumptions of its research investigation.

c: PHASE PLANNING

1) Examine the nature of the dataset and its associated meta-data if available, e.g: data types, structure, size, and format.
2) Verify data understanding by explaining the meanings of the attributes and the scope of their values.
3) Investigate the quality of the data, e.g: missing values, outliers.
4) Prospect the potential of the data based on the defined process objectives and using statistical tests, measurements, and other methods .
5) Gain insight into the data trends, distribution, and relationships between its features.
6) Confirm the relevance, sufficiency, and adequacy of data to fulfil the defined process objectives.

d: PHASE PERFORMING

1) Perform and record the planned data exploration activities.
2) Justify and record the chosen phase options and alternatives.
3) Document the problems, gaps, and limitations that are encountered when carrying out the planned phase activities.

e: PHASE VALIDATION

1) Validate that all the exploration procedures and activities carried out in the phase have performed the following activities:
   a) Investigated the targeted data in terms of their nature and quality;
   b) Understood the metabolomics data in terms of their comprehensibility, correctness and interpretability by describing the data in a language which bridge the gap between data miner and domain expert background and terminologies;
   c) Prospected the data comprehensively and viewed the it from the required perspectives;
2) Validate that the exploration activities carried out were comprehensive and thorough and covered the entire preprocessed dataset and its associated meta-data if available and encompassed all their related aspects.
3) Validate that the exploration activities have been carried out in the light of the defined process objectives and that in the case of hypothesis-driven data mining, were also in line with the aims of metabolomics study and with the goals, hypothesis and assumptions of its research investigation.
4) Validate that data was checked in its relevance, sufficiency, and adequacy to fulfil the defined process objectives.
5) Validate that all the data exploration activities have been carried out, recorded, and justified according to the plan.

6) Validate that all the problems, gaps and limitations encountered in the phase activities have been recorded according to the plan.

f: PHASE REPORTING

1) Define the applicable reporting standards, based on the available standards in data mining and metabolomics.
2) Report data exploration results including data investigation, data understanding and prospecting.
3) Report the procedures carried out and as well as their involved justification, in the case of deciding alternatives and options.
4) Report the phase validation outcomes.
5) Report the human interactivity as this phase is shared by both data miner and domain expert. This is done by reporting the performer of each of the phase performed activities.
6) Report the problems, gaps, and limitations, which might be encountered during the performance of the phase activities.

4) TECHNIQUE SELECTION

This phase provides a strategy for selecting and justifying the selection of a data mining technique that should achieve the process objectives and suit the targeted data. The strategy considers the requirements and feasibility of the selected technique and defines its performance measurability and success criteria. The results of this phase include the selection of the technique and its and justification, as well as the various factors, considerations and assessments involved.

a: PHASE PREREQUISITES

- The preprocessed metabolomics dataset and its associated meta-data if available.
- The data exploration report including the results of data investigation, understanding and prospecting.
- The defined process objectives.
- The aims of the study and their relationship with the goals of the metabolomics investigation and its hypotheses and assumptions.
- Background information/knowledge regarding the data mining, including: its approaches, goals, tasks and techniques.
- Background information/knowledge regarding the requirements of data mining techniques application, including: time, cost, and expertise as well as the availability of the software and hardware infrastructure.

b: PHASE OBJECTIVES

- Selecting the data mining technique(s) that fulfil the defined process objectives and suit the targeted metabolomics data.
- The selected technique performance must be measurable in model evaluation.
- The selected technique must have the potential to achieve the defined process objective.
- The application of the selected technique must be feasible, and it must consider the process management constraints and available resources.
- The selection procedures must be comprehensive and unbiased, and it must cover all possible data mining techniques.
- The selection strategy must be flexible and adjustable and consider the selection of alternative techniques in the case of feedback or iteration.
- The selection of white-box data mining techniques must be encouraged over the selection of the black-box ones, as they allow more justification, reasoning, and explanation of the modelling results.
- The selection must consider the possibility of using more than one data mining technique.

c: PHASE PLANNING

1) Identify the suitable data mining approach to be used for model building, based on the type of the process objectives defined in phase 1 and in the light of data mining approaches.
2) Match the process objective to the data mining goals and tasks.
3) Match the process objective to the available data mining techniques and take into consideration the results of the data exploration phase regarding the nature, quality, and potential of the targeted data.
4) Based on the steps 1-3, select the data mining technique that would fulfil the defined objectives and suit the targeted data.
5) Identify the resources, which are required for applying the selected technique, e.g. software, hardware, expertise, etc.
6) Perform the following assessments on the candidate technique:
   a) Assess the potential fulfilment of the defined process objectives by the candidate technique;
   b) Assess the suitability of the candidate technique to the nature and quality of the data, as well as to its trends and expected patterns, based on the results of the phase data exploration phase;
   c) Assess the intensity of the acclimatisation activities, which are required by the candidate technique;
   d) Assess the availability of the resources required for applying the candidate technique;
   e) Assess the feasibility of the application of the candidate technique in terms of cost and effort;
7) In the case of the candidate techniques failure in assessments in step 6, consider selecting an alternative technique by repeating the steps 4-6.
8) Identify the performance measurements, which are applicable to the selected technique, e.g. accuracy, sensitivity, precision, specificity, etc.
9) Define the success criteria, based on the identity measurement.

10) Define a mechanism for applying the selected modelling technique, considering its required resources identified in 5 and assessed in 6.

d: PHASE PERFORMING

1) Perform and record the planned modelling technique selection activities.
2) Justify and record the chosen phase options and alternatives.
3) Document the problems, gaps, and limitations that are encountered when carrying out the planned phase activities.

e: PHASE VALIDATION

1) Validate that data mining technique selection has been performed comprehensively in an unbiased way according to the defined selection criteria and the identified constraints.
2) Validate that the selected modelling technique is suitable for fulfilling the defined process objectives and it suits the targeted data.
3) Validate that an appropriate and sufficient justification has been given for the selection of the technique.
4) Validate that all the activities in the phase have been carried out, recorded and justified according to the plan.
5) Validate that all the problems, gaps and limitations encountered in the phase activities have been recorded according to the plan.

f: PHASE REPORTING

1) Define the applicable reporting standards, based on the available standards in data mining and metabolomics.
2) Report the selected data mining technique(s) and its/their implementation requirements, including: software tools, hardware, and expertise as well as other necessary information regarding the model inputs, parameters and expected outputs, as well as the constraints of their application.
3) Report the procedures carried out and as well as their involved justification, in the case of deciding alternatives and options.
4) Report the phase validation outcomes.
5) Report the human interactivity as this phase is shared by both data miner and domain expert. This is done by reporting the performer of each of the phase performed activities.
6) Report the problems, gaps, and limitations, which might be encountered during the performance of the phase activities.

5) DATA ACCLIMATISATION

Data acclimatisation involves processing and preparing the data to suit the needs and requirements of the selected data mining technique(s) which are performed in the light of the defined process objectives. This phase typically generates two or more datasets, which are used for model building, training, validation and testing.

a: PHASE PREREQUISITES

- The preprocessed metabolomics dataset and its associated meta-data if available.
- The defined process objectives.
- The data exploration report, including: the results of data investigation, understanding and prospecting.
- The selected data mining technique(s), as chosen and reported in the previous phase.
- The requirements of the implementation of the selected technique(s) including: software tools, hardware, and expertise, modelling inputs format, parameters, and expected outputs, as well as their application constraints, including: time, cost, and expertise as well as the availability of the software and hardware infrastructure.
- Background knowledge/information regarding the basic concepts of data processing in computing, including: databases, data types, data structures, file, and data format.
- The process quality assurance policy.
- The process standards.

b: PHASE OBJECTIVES

- Acclimatising the targeted data to suit the modelling technique(s) and meet its/ their modelling requirements in terms of their input data structure, data types, file format, size and quality.
- Preparing the metabolomics data for model building and training, as well as for model testing and evaluation.
- The data acclimatisation procedures must be carried out in the light of the defined process objectives and to fulfil those precise objectives intended by the selected data mining modelling technique(s).
- The requirements of data acclimatisation procedures must be considered including: time, cost, expertise as well as software and hardware infrastructures.
- The data acclimatisation procedures must be adequate, sufficient and comprehensive to cover all the data required for model building, training and testing.
- The data acclimatisation procedures must not influence the data through twisting, drifting, or changing their meanings, and must not cause problems, gaps, or loses of information.
- Acclimatisation must avoid the issues which may lead to model over-fitting or under-fitting.
- Acclimatisation should consider selecting more than one technique (compound modelling) which involves sub-modelling, e.g. PCA-ANN. Sub-modelling should be performed iteratively within the boundaries and context of data acclimatisation.

c: PHASE PLANNING

1) Identify the requirements for the selected data mining technique in terms of the data structure, data types, file format, data size, and quality.

2) Identify the data acclimatisation procedures, which must be applied to the targeted data, in order to make it suitable for the selected data mining technique.
3) Identify the requirements of the data acclimatisation procedures.
4) Perform the identified acclimatisation procedures and record their purpose and description.
5) In the case of compound modelling, e.g. PCA-ANN, the following steps should be carried out for each of the sub-models involved:
    a) The data is acclimatised for the first sub-modelling technique in the sequence, e.g. PCA;
    b) The sub-model is built and evaluated;
    c) The output of the sub-modelling are forwarded towards the next phase (model building) after performing the necessary procedure and activities to acclimatise and prepare the data to suit the requirement of the main model and achieve its objectives (In the case of multiple sub-modelling, the last model in the sequence is treated as the main model, and the steps a-c is repeated for each of the sub-models);
6) Decide the data splitting strategy which will be used to choose the model testing data, e.g. holdout, bootstrap, random sub-sampling, validation and cross-validation.
7) Split the data according to the strategy decided to obtain the data which will be used in model testing.

d: PHASE PERFORMING

1) Perform and record all the planned activities which make the data suitable for data mining through model building, training and testing.
2) Justify and record the chosen phase options and alternatives.
3) Document the problems, gaps, and limitations that are encountered when carrying out the planned phase activities.

e: PHASE VALIDATION

1) Validate that the data have been acclimatised in the light of the modelling objectives.
2) Validate that the data have been acclimatised according to the identified requirements for the selected data mining technique
3) Validate that acclimatisation considered the data exploration report.
4) Validate that any sub-modelling involved in data acclimatisation has been built according to the iterative cycle of acclimatisation, modelling, and evaluation, within the boundaries of the acclimatisation phase, and taking into consideration its integration with the main model in the process.
5) Validate that all the activities in the phase have been carried out, recorded and justified according to the plan.
6) Validate that all the problems, gaps and limitations encountered in the phase activities have been recorded according to the plan.

f: PHASE REPORTING

1) Define the applicable reporting standards, based on the available standards in data mining and metabolomics.
2) Report and retain the acclimatised data.
3) Report the procedures carried out and as well as their involved justification, in the case of deciding alternatives and options.
4) Report the phase validation outcomes.
5) Report the human interactivity as this phase is shared by both data miner and domain expert. This is done by reporting the performer of each of the phase performed activities.
6) Report the problems, gaps, and limitations, which might be encountered during the performance of the phase activities.

6) MODEL BUILDING

This phase involves building and training a data mining model, that both fulfills the defined process objectives and suits the targeted dataset(s) that is/are acclimatised in the previous phase by applying the selected data mining modelling technique(s).

a: PHASE PREREQUISITES

- The acclimatised metabolomics dataset allocated for model building and training and its associated meta-data if available.
- The selected data mining technique.
- The defined process objectives.
- The requirements of the selected data mining technique implementation, including: software tools, hardware and expertise as well as other necessary information regarding the model inputs, parameters, and expected outputs, as well as the constraints of their application.
- The process standards.

b: PHASE OBJECTIVES

- Building a data mining model using the metabolomics data acclimatised in the previous phase to fulfil the defined process objectives.
- Training the model using the acclimatised data to reach its maturity level but at the same time avoiding the over-fitting of the model.
- The model must fulfil the defined process objectives.
- In the case of hypothesis driven data mining, the model must also be in line with the aims of the metabolomics study experiment hypothesis, assumption and research question. If not, the model should provide the sufficient justification, reasoning or explanation of its results.
- The data mining model must be measurable and testable through model evaluation.

- The data mining model must be flexible and adjustable through changing its settings and parameters.
- The model must consider the requirements of the later phases, including: model evaluation, knowledge presentation and evaluation, and results deployment.
- Training must avoid model over-fitting or under-fitting.

c: PHASE PLANNING

1) Assign the resources required for model building and training, which was identified in phase 4.
2) Build a data mining model with the technique selected in phase 4 and using the data split acclimatised in phase 5, which was allocated for model building and training.
3) Ensure the flexibility of the model by allowing the adjustment of its parameters.
4) Define a mechanism for delivering the model taking into consideration model reporting standards, e.g. PMML and requirements of other phases, including: model evaluation, knowledge presentation, knowledge evaluation, and results deployment.

d: PHASE PERFORMING

1) Perform and record all the planned activities involved in model building and training.
2) Justify and record the chosen phase options and alternatives.
3) Document the problems, gaps, and limitations that are encountered when carrying out the planned phase activities.

e: PHASE VALIDATION

1) Validate that the built data mining model achieved the defined process objectives using the acclimatised data.
2) Validate that the model has been trained with the acclimatised training data.
3) Validate the model measurability, testability, adjustment, flexibility
4) Validate that the requirements of the model evaluation, presentation, and delivery as knowledge in the later phases have been considered.
5) Validate that the model has been delivered according to the defined delivery plan in the appropriate format, e.g. PMML.
6) Validate that all the activities in the phase have been carried out, recorded and justified according to the plan.
7) Validate that all the problems, gaps and limitations encountered in the phase activities have been recorded according to the plan.

f: PHASE REPORTING

1) Define the applicable reporting standards, based on the available standards in data mining and metabolomics.
2) Report the final data mining model using the delivery mechanism defined in the phase plan.
3) Report the model requirements, including: software tools, hardware and expertise, modelling inputs format, parameters and expected outputs, as well as their applying constraints.
4) Report the procedures carried out and as well as their involved justification, in the case of deciding alternatives and options.
5) Report the phase validation outcomes.
6) Report the human interactivity as this phase is shared by both data miner and domain expert. This is done by reporting the performer of each of the phase performed activities.
7) Report the problems, gaps, and limitations, which might be encountered during the performance of the phase activities.

7) MODEL EVALUATION

This phase involves testing and validating the data mining model, based on the defined process objectives and using the applied technique measurability criteria. Model evaluation is usually performed using a separated data split which must be allocated for model validation during data acclimatisation phase.

a: PHASE PREREQUISITES

- The acclimatised metabolomics dataset allocated for model testing and its associated meta-data if available.
- The data mining model.
- The data mining model requirements, including: software tools, hardware and expertise as well as other necessary information regarding the model inputs, parameters, and expected outputs, as well as the constraints of their application.
- The defined process objectives.
- Background knowledge/information regarding the model evaluation.
- The process quality assurance policy.

b: PHASE OBJECTIVES

- Evaluating the data mining model validity using the decided testing data split.
- Evaluating the data mining model achievement of the defined process objectives such as the data-driven objectives (e.g. discover interesting or hidden patterns, trends, relationships in the data) or hypothesis driven objectives (e.g. classify samples into predefined classes or according to their natural occurrence classes).
- Evaluating the data mining modelling success criteria defined in both objectives definition and modelling techniques selection phases.
- Evaluating the data mining model performance according to its defined measurements.
- Performance measurements must be correct, sufficient, relevant and applicable to the data mining model.
- The data mining model should be neither over-fitted nor under-fitted.
- Model evaluation activities and procedures should be correct, comprehensive and unbiased.

- Model evaluation should be repeated each time a new data mining model is built or when the existing model is changed, adjusted or recreated.

c: PHASE PLANNING

1) Test the data mining model, which was built in phase 6 using the data acclimatised in phase 5 and allocated for model testing.
2) Measure the model performance according to the criteria defined in phase 4, e.g. accuracy, sensitivity, specificity, scalability, etc.
3) Validate the model fitting (over-fitness or under-fitness).
4) Assess the model fulfilment of the defined process objectives.
5) Assess the model according to the success criteria defined in the process objective, e.g. classify samples, uncovering interesting/ hidden patterns, trends, or relationships in the data.
6) Assess the model according to the success criteria defined in the technique selection phase.
7) In the case of obtaining unsatisfactory results from the steps 1-6, consider selecting a different data mining technique and building an alternative model or a combination of models. This can be done by performing feedback to technique selection phase, and re-running the data acclimatisation, model building, and model evaluation phases.

d: PHASE PERFORMING

1) Perform and record all the planned activities involved in model evaluation.
2) Justify and record the chosen phase options and alternatives.
3) Document the problems, gaps, and limitations that are encountered when carrying out the planned phase activities.

e: PHASE VALIDATION

1) Validate that the model validity was tested using the selected data split.
2) Validate that the model performance was measured using the sufficient, correct and relevant measurements.
3) Validate that the data mining model has fulfilled the objectives defined in the first phase.
4) Validate that the data mining model has passed the success criteria defined in the objectives definition phase.
5) Validate that the model is neither over fitted nor under fitted.
6) Validate that all the activities in the phase have been carried out, recorded and justified according to the plan.
7) Validate that all the problems, gaps and limitations encountered in the phase activities have been recorded according to the plan.

f: PHASE REPORTING

1) Define the applicable reporting standards, based on the available standards in data mining and metabolomics.
2) Report the model evaluation outcomes, including: the results of model testing, performance measurements, success assessment, over fitness and under fitness checks.
3) Report the procedures carried out and as well as their involved justification, in the case of deciding alternatives and options.
4) Report the phase validation outcomes.
5) Report the human interactivity as this phase is shared by both data miner and domain expert. This is done by reporting the performer of each of the phase performed activities.
6) Report the problems, gaps, and limitations, which might be encountered during the performance of the phase activities.

8) KNOWLEDGE PRESENTATION

This phase involves presenting the model built and validated in the previous phase in a form which presents the acquired metabolomics knowledge. Knowledge presentation might require performing complex visualisation techniques in order to facilitate interactive presentation of knowledge.

a: PHASE PREREQUISITES

- The data mining model.
- The data mining model evaluation.
- The aims of the metabolomics study and its investigation goals, hypotheses, and assumptions.
- The domain expert requirements regarding the presentation knowledge.
- Background knowledge/information regarding knowledge presentation recommended facilities, features, and best practices.
- Background knowledge/information regarding the available knowledge presentation methods, techniques and tools.
- The requirements of knowledge presentation tools, including: software tools, hardware and expertise as well as other necessary information regarding the model inputs, parameters and expected outputs, as well as the constraints of their application.
- The process standards.

b: PHASE OBJECTIVES

- Presenting the discovered knowledge according to the requirements and preferences of the domain expert.
- The presented Knowledge must be interpretable and comprehensible by the domain expert.
- Knowledge presentation must support human interaction, e.g. the adjustment of the model inputs and parameters, and evaluating it impacts on the outcomes of the model.

- Knowledge presentation must support traceability of model inputs, parameters, justifications, decisions and outcomes.
- Knowledge presentation should be performed according to metabolomics domain expert perspective, as the domain expert is the principle knowledge presentation stakeholder.
- Knowledge presentation must allow multiple presentations of the same model that reflects the knowledge from different perspectives.
- Knowledge presentation must encourage visualisation where possible.

c: PHASE PLANNING

1) Review the knowledge presentation features required for metabolomics data mining.
2) Identify the requirements and preferences of the domain expert regarding the knowledge presentation.
3) Match the defined knowledge presentation features and facilities in step 1 and the identified domain expert requirements in step 2 to the available knowledge presentation techniques and tools.
4) Assess the requirements of the decided knowledge presentation technique, including: software tools, hardware and expertise as well as inputs, parameters, and expected outputs, and application constraints, including: time, cost, and expertise as well as the availability of the software and hardware infrastructure.
5) Prototype the knowledge presentation iteratively through the steps 3-4 until the desired features are satisfied, the requirements of the domain expert are met, and the phase objectives are achieved.
6) Evaluate the knowledge presentation interpretability as a metabolomics knowledge (to be carried out by domain experts).
7) Present the successful knowledge presentation as a metabolomics knowledge.

d: PHASE PERFORMING

1) Perform and record all the planned activities involved in model building and training.
2) Justify and record the chosen phase options and alternatives.
3) Document the problems, gaps, and limitations that are encountered when carrying out the planned phase activities.

e: PHASE VALIDATION

1) Validate that the knowledge presentation has met the domain expert identified requirements.
2) Validate that the presented knowledge:
    a) Is interpretable and comprehensible by the domain expert as metabolomics knowledge.
    b) Supports of human interaction, e.g. adjusting model inputs and parameters and evaluating their impact on the model outcomes and results.
    c) Supports traceability of model inputs, parameters, justifications, decisions and outcomes.
    d) Supports multiple perspectives.
3) Validate that all the activities in the phase have been carried out, recorded and justified according to the plan.
4) Validate that all the problems, gaps and limitations encountered in the phase activities have been recorded according to the plan.

f: PHASE REPORTING

1) Define the applicable reporting standards, based on the available standards in data mining and metabolomics.
2) Report metabolomics presented knowledge.
3) Report the procedures carried out and as well as their involved justification, in the case of deciding alternatives and options.
4) Report the phase validation outcomes.
5) Report the human interactivity as this phase is shared by both data miner and domain expert. This is done by reporting the performer of each of the phase performed activities.
6) Report the problems, gaps, and limitations, which might be encountered during the performance of the phase activities.

9) KNOWLEDGE EVALUATION

In this phase, the knowledge acquired and presented earlier is evaluated from a metabolomics perspective. This is performed in terms of its fulfillment of the objectives defined in the first phase, as well as in terms of its validity as a metabolomics knowledge, based on the background knowledge.

a: PHASE PREREQUISITES

- The presented knowledge.
- The defined process objectives
- The aims of the metabolomics study and its investigation goals, hypotheses, and assumptions.
- Background knowledge/information regarding the data mining approaches, goals and tasks, and their modelling techniques objectives.
- Domain expert background knowledge and experience
- Metabolomics biochemical principles.
- The results reported or published in the metabolomics literature which is relevant to the metabolomics research investigation, study and assay.
- The process quality assurance policy.

b: PHASE OBJECTIVES

- Evaluating the acquired knowledge fulfilment of the defined process objectives and its achievement of the aims of the study, the goals of the metabolomics

investigation and its compliance with the research hypothesis and assumption.
- Evaluating the acquired knowledge against the results reported or published in the metabolomics literature which is relevant to the metabolomics research investigation, study and assay.
- Validating the acquired knowledge according domain expert background experience and knowledge, based on biochemical principles and commonsense.
- Identify possible investigation questions or hypotheses propagated by the acquired knowledge.
- Knowledge evaluation activities which involve human judgments must be unbiased, traceable and reproducible. It must be justified and backed with scientific evidence and the literature where possible.
- Knowledge evaluation must be performed from metabolomics and biochemical perspectives.

c: PHASE PLANNING

1) Identify the modelling objectives related to the acquired knowledge.
2) Match the modelling objectives to its related aspects of acquired knowledge.
3) Decide whether the acquired knowledge has fulfilled the defined process objectives or has not.
4) In the case of hypothesis driven data mining, evaluate the acquired knowledge achievement of the aims of metabolomics study and its relationship with the goals of the investigation, the research hypothesis and assumption.
    a) Identify the aims of the metabolomics study related to the acquired knowledge;
    b) Match the aims of the metabolomics study to its related aspect of acquired knowledge;
    c) Decide whether the acquired knowledge have achieved the aims of metabolomics study or have not;
    d) Assess the acquired knowledge consistency with the goals of the original research investigation;
    e) Assess the acquired knowledge compliance with the metabolomics research hypothesis and assumption;
5) Evaluate the acquired knowledge against the results reported or published in the metabolomics literature, which are relevant to the metabolomics research investigation, study and assay.
    a) Identify the metabolomics results related to the acquired knowledge (through researching the metabolomics literature and technical reports);
    b) Compare and contrast the acquired knowledge with the relevant results reported or published in the metabolomics literature;
    c) Assess the acquired knowledge consistency and compliance with the relevant results reported or published in the metabolomics literature;
6) Validate the acquired knowledge according to the metabolomics biological and biochemical principles, the background knowledge, and commonsense.
7) Validate the acquired knowledge according to the domain expert knowledge and experience.
8) Identify any possible investigation questions or hypotheses propagated by the acquired knowledge.
9) Recommend a mechanism for answering the propagated knowledge questions, e.g. defining new objectives for a further iteration in the process, altering the current objectives, or through process feedback, recommendation for further experiment or process execution.

d: PHASE PERFORMING

1) Perform and record all the planned activities involved in knowledge evaluation.
2) Justify and record the chosen phase options and alternatives.
3) Document the problems, gaps, and limitations that are encountered when carrying out the planned phase activities.

e: PHASE VALIDATION

1) Validate acquired knowledge fulfilment of the defined process objectives.
2) Validate the acquired knowledge achievement of the aims of metabolomics study and its relationship with the goals of the investigation and the research hypothesis and assumption.
3) Validate the acquired knowledge consistency and compliance with the relevant results reported or published in the metabolomics literatures which are relevant to the metabolomics research investigation, study and assay.
4) Validate that the acquired knowledge was checked according to the metabolomics biological and biochemical principles, the background knowledge, and commonsense.
5) Validate that the acquired knowledge was checked according to the domain expert background knowledge and experience.
6) Validate that possible investigation questions or hypotheses propagated by the acquired knowledge have been identified.
7) Validate that knowledge evaluation have been performed by the domain expert from metabolomics and biochemical perspectives.
8) Validate that human related judgments and decisions were justified and backed up with evidences where possible, e.g. the scientific literature.
9) Validate that knowledge evaluation was unbiased, justifiable, traceable and reproducible where possible.

10) Validate that all the activities in the phase have been carried out, recorded and justified according to the plan.
11) Validate that all the problems, gaps and limitations encountered in the phase activities have been recorded according to the plan.

f: PHASE REPORTING

1) Define the applicable reporting standards, based on the available standards in data mining and metabolomics.
2) Report the model evaluation results.
3) Report the literature results and publications which acquired knowledge have been evaluated against.
4) Report human related judgments and decisions which must be justified and backed up with evidence and the scientific literature where possible.
5) Report the investigation questions and hypothesis propagated by the acquired knowledge.
6) Report the procedures carried out and as well as their involved justification, in the case of deciding alternatives and options.
7) Report the phase validation outcomes.
8) Report the human interactivity as this phase is shared by both data miner and domain expert. This is done by reporting the performer of each of the phase performed activities.
9) Report the problems, gaps, and limitations, which might be encountered during the performance of the phase activities.

10) DEPLOYMENT

This phase aims to deploy the acquired knowledge through a mechanism that enables effective knowledge utilisation. It involves selecting appropriate deployment mechanisms in the light of the defined process objectives and within the available resources, as well as the selection of the particular deliveries which must be deployed with the knowledge.

a: PHASE PREREQUISITES

- Background knowledge/information regarding the available deployment options, mechanisms, techniques and implementation requirements, including: software tools, hardware and expertise as well as other necessary information regarding the model inputs, parameters and expected outputs, as well as the constraints of their application.
- The aims of the metabolomics study and its investigation goals, hypotheses, and assumptions.
- The defined process objectives
- The justification of data technique selection.
- The acclimatised data.
- The data mining model.
- The model evaluation report.
- The presented knowledge.
- Knowledge evaluation.
- The process standards.

b: PHASE OBJECTIVES

- Defining a mechanism of delivering, retaining, accessing and utilising the acquired data mining knowledge for further analysis.
- Deploying the acquired knowledge and its supplement deliveries, either as intermediate results of the process iteration or as final results.
- Knowledge deployment should be carried out according to the requirements of domain experts and other stakeholders if there are any.
- Knowledge deployment must consider the availability and requirements of the possible knowledge delivery techniques, e.g. software tools, databases, hardware and expertise, as well as their constraints.
- Knowledge deployment should take into consideration knowledge utilisation and possible future analysis.
- Knowledge must be based on the requirements and preferences of domain experts (biologist).
- Knowledge deployment should be carried out as a joint phase between data miner and domain expert (biologist).

c: PHASE PLANNING

1) Decide whether the results to be deployed as intermediates of a process iteration or as final results of a process termination.
2) Identify the deliveries need to be reported as supplements of the acquired knowledge.
3) Identify the requirements of delivering; retaining, accessing and utilising the acquired knowledge, e.g. file formats, user interfaces, system interfaces, required software tools, databases, etc.
4) Define a mechanism of delivering, retaining, accessing, and utilising the acquired knowledge and its supplement deliveries.
5) Assess the requirements and constraints of implementing the deployment mechanism.
6) Define a mechanism of deploying the acquired knowledge and its supplement deliveries, e.g. DBMS, software, web, based or embedded system.
7) Deploy the data acquired knowledge and its associated supplementary deliveries.

d: PHASE PERFORMING

1) Perform and record all the planned deployment activities.
2) Justify and record the chosen phase options and alternatives.
3) Document the problems, gaps, and limitations that are encountered when carrying out the planned phase activities.

e: PHASE VALIDATION

1) Validate that the acquired knowledge and its associated supplementary deliveries have been deployed, retained, and provided with facilities for accessing and utilising

the knowledge according to the defined mechanism and the identified requirements.
2) Validate that all the activities in the phase have been carried out, recorded and justified according to the plan.
3) Validate that all the problems, gaps and limitations encountered in the phase activities have been recorded.

f: PHASE REPORTING

1) Report the deployed knowledge and its associated deliveries, based on the defined requirements and mechanism of delivering; retaining, accessing and utilising the knowledge.
2) Report the procedures carried out and as well as their involved justification, in the case of deciding alternatives and options.
3) Report the phase validation outcomes.
4) Report the human interactivity as this phase is shared by both data miner and domain expert. This is done by reporting the performer of each of the phase performed activities.
5) Report the problems, gaps, and limitations, which might be encountered during the performance of the phase activities.

11) PROCESS EVALUATION

This phase concerns evaluating the execution of the process model in terms of the flow of its phases and the validity of the tasks applied within the performed phases. It also ensures the quality of the process deliveries through the defined mechanisms of cross-deliveries validation.

a: PHASE PREREQUISITES

- The human interaction aspects related to the process execution.
- The project management aspects related to the process execution.
- The process quality assurance policy.
- The process standards.
- The process model layout design regarding its phases, inputs, deliveries, flow, iteration, and feedback.
- The defined process objectives
- The justification of data mining technique selection.
- The acclimatised data.
- The data mining model.
- The model evaluation report.
- The presented knowledge.

b: PHASE OBJECTIVES

- Evaluating the execution of the process in terms of its compliance with the defined flow, feedback and iteration, as well as the inputs and outputs of the process phase.
- Evaluating the process phases running in terms of their compliance to the process phases objectives, planning, activities, validation and reporting.
- Evaluating the process cross-deliveries validation as defined in the process.
- Evaluating the process compliance of management, quality, and human interaction and standardisation issues throughout the process.
- Deciding the process termination, iteration or performing limited feedback.

c: PHASE PLANNING

1) Evaluate the process execution in terms of:
   a) The flow between phases during the process normal execution, iteration,and feedback;
   b) The feedback between phases was carried out according to the process model defined mechanism;
   c) The process iterations was performed according to the process model defined mechanism;
2) Evaluate that the phases of the process were run according to the defined cycle, including: objectives, planning, activities, and validation and reporting.
3) Perform the cross-deliveries validation, which checks the validity and consistency of the deliveries of earlier phases with those of the later phases.
4) Evaluate the process in terms of its consideration of issues including:
   a) Project management issues consideration on the level of the process and its subsequent phases, e.g. tasks partitioning, planning, etc;
   b) Quality assurance issues regarding the process inputs, deliveries and procedures, either on the level of the process or on the level of its subsequent phases;
   c) Human interaction issues on the level of the process and its subsequent phases;
   d) Standardisation consideration throughout the process and its subsequent phases, including: data, report, or procedural standardisation;
5) Based on the results of 1-4 and the results of knowledge evaluation, decide, either to terminate the process, iterate throughout the process or perform limited feedback.

d: PHASE PERFORMING

1) Perform and record all the planned activities involved in process evaluation.
2) Justify and record the chosen phase options and alternatives.
3) Document the problems, gaps, and limitations that are encountered when carrying out the planned phase activities.

e: PHASE VALIDATION

1) Validate that the process execution has been evaluated, based on the process model flow, feedback and iterations.

2) Validate that the evaluation covered all the phases in the process.
3) Validate that the process cross-delivery validation has been carried out according to the process model.
4) Validate that the process considered the supplementary issues including: management, quality, human interaction and standardisation.
5) Validate that all the activities in the phase have been carried out, recorded and justified according to the plan.
6) Validate that all the problems, gaps and limitations encountered in the phase activities have been recorded.

f: PHASE REPORTING

1) Report the results of process evaluation, including: process execution, phase running, cross-delivery validation and its consideration of issues, including: management, quality, human interaction and standardisation.
2) Report the decision regarding process termination, iteration or feedback.
3) Report the procedures carried out and their justification when options and alternatives were decided.
4) Report the phase validation outcomes.
5) Report the human interactivity as this phase is shared by both data miner and domain expert. This is done by reporting the performer of each of the phase performed activities.
6) Report the problems, gaps, and limitations, which might be encountered during the performance of the phase activities.

E. PROCESS MODEL SUPPLEMENTARY SUPPORT

The process model also integrates its support for data mining practical aspects by embedding them in its layout design and in the description of its phases. Human interaction is considered in all phases, either on the level of process phases execution or on the level of the tasks within its phases. The process uses a colour coding system to represent the participant in each phase, e.g. data miner, domain expert or both. The process structure and description also provides details regarding who is doing what, where, and how. Management issues are considered on the level of the process, as well as on the level of its phases. The objectives of the process are defined in the first phase in terms of their measurability, success criteria, and feasibility. These are used later for technique selection, and for the evaluation of the discovered knowledge and process execution. On the other hand, management aspects are in structuring the tasks within the process phases, where the objectives of the phase are set, and its planned activities are executed, validated and reported. Quality assurance is embedded in the design of the process model and also considered in its execution. The process model validates the quality of the data, as well as the model and its presented knowledge. The execution of the process model is also validated, as well as activities within its phases. The process model also provides an extra level of validation on the level of its deliveries. Metabolomics and data mining standards were both used for designing the process layout, as well as their available ontologies. The process also encourages using the reporting standards in reporting its deliveries of its phase, e.g. PMML.

F. CROSS-DELIVERY VALIDATION

The process model defines a mechanism for evaluating process deliveries in terms of backward report-level consistency check. Cross-delivery validation was inspired by software engineering v-model. It aims to provide a high level quality assurance for the process on the level of its generated deliveries. The results of cross-delivery validation can cause feedback to resolve the inconsistencies, as well as process iteration. However, any feedback should be executed according to the process flow, where all the phases between the two deliveries must be re-visited to ensure that they are all aware of the changes made to their processor phases. Cross-validation in the process model, covers the relationship between the following deliveries:

- ***Modelling Objectives vs. Knowledge Evaluation:*** this validates the modelling objectives in light of the outcomes of knowledge evaluation regarding their compatibility with the aims of metabolomics study and the original investigation question and assumptions. It validates the testability of the investigation hypothesis,; It validates the achievable objectives using the available metabolomics data which must be sufficient, adequate and relevant in fulfilling the aims of the metabolomics study in discovering new knowledge or testing its implied hypothesis. It validates measurability and testability, realistic achievability within the given time, cost, human expertise, hardware and soft- ware parameters.
- ***Modelling Objectives vs. Presented Knowledge:*** this validates the defined objectives in light of presented knowledge. The presented knowledge may uncover flaws in the defined objectives regarding their achievability or any unrealistic requirements.
- ***Modelling Objectives vs. Selection Justification:*** this validates the defined objectives against the outcome of the modelling techniques selection phase. The justification may uncover flaws in the defined objectives regarding their achievability within the possible data mining techniques, or the feasibility of its application within the available resources and project constraints, time, cost, and the availability of human expertise and software tools.
- ***Selection Justification vs. Data Mining Model:*** this validates the justification for modelling techniques in light of the outcomes of model building. A bad model might reflect flaws in the data mining technique selection process. The model might uncover problems with justifying the selection with regards to the ability of the model to fulfill the defined objectives or its compatibility with the data, or its ability to justify a sufficient and appropriate selection; or with the feasibility assessment

of the applicability of the model within the available resources and project constraints, time, cost, and the availability of human expertise and software tools.

- ***Selection Justification vs. Presented Knowledge:*** validate the justification for modelling technique selection in light of the presented knowledge. The presented knowledge might uncover flaws regarding the selected modelling technique and its associated justification regarding its foreseen achievability of the modelling objectives and aims of metabolomics study; interpretability into biological knowledge; and its presentability in a form of metabolomics or biological knowledge.
- ***Acclimatised Data vs. Data Mining Model:*** this validates the acclimatised data in light of the model building. Problems with the built model or with its quality may uncover gaps or flaws in the fulfilment of the requirements in the modelling techniques used for acclimatised data, be they data types, structure and format, volume or quality, e.g. bad handling of missing values, outliers, data drift, over-fitting, lost of information, problems in sampling, data merging, and conversions, etc.
- ***Data Mining Model vs. Presented Knowledge:*** validate the data mining model in light of the presented knowledge. The presented knowledge might uncover flaws in the built model regarding its success in uncovering, reflecting and presenting the data mining knowledge hidden or interesting patterns, trends, or association; the model validating and from a metabolomics requirements perspective and its suitability to be presented as metabolomics biological knowledge and might imply a second hand problems regarding its variables traceability, selection justifiability or decisions reproducibility which might uncover further problems in the data it is applied to regarding its adequacy, sufficiency, relevance, pre-processing and data mining procedures or in the objectives it is trying to achieve.

G. MAPPING OF THE PROCESS MODEL

The process model utilises the concept of mapping where the process model is broken down into groups of stages which consist of a set of the generic tasks that are then mapped to a specific model that is customised to achieve the specific aims of the metabolomics study and the specific objectives of the data mining approach adopted. The customisation must meet the particular requirements of metabolomics study and the particular needs of the data generated by the assays. Figure 13 provides an illustration of the process concept of mapping as adopted by the proposed process model.

The Generic Model is mapped to a specific model which is customised to achieve the particular aims of the metabolomics study and fulfil the objectives of the data mining approach adopted, as well as meeting the requirements of its metabolomics study and the needs of the data generated by the assays. It can be customised and can then be applied to all metabolomics applications regardless of the metabolic approach and data acquisition techniques or the applied data mining techniques. The specific model maps the generic process model into a more specific one.

The Specific Model The specific model maps the generic process model into a more specific one. The specific model takes into consideration all factors and issues specific to a particular type of metabolomics study which might influence the applicability of the process and definition of its objectives and the selection of its applied data mining technique(s).

Implementation Model The implementation model is a realisation of the process model into an applicable one. It realises the process phases and activities into applicable functions and algorithms that can be applied to a specific metabolomics data. The relationships of the various models are shown in Fig 5.1 using UML notation.

The Process Instance Process instance is an instance of the abstract generic process model created in the form of an implementation model that is ready to be applied to the target data to acquire the desired knowledge.

PROCESS MODEL INSTANTIATION AND EXECUTION

The **process execution** is carried out by running the internal tasks of the process phases and performing their planned activities in order to generate their defined outcomes and deliverables either as part of the normal process flow; or as part of a feedback or a process iteration. **The Normal process flow** involves executing the process phases in order as shown in Figure 12, while **process iteration** involves the repeated execution of all phases maintaining the normal flow of their execution and taking into consideration the inputs and deliveries of the phase execution in each new iteration. An iteration might be triggered by a significant change in process objectives; to formulate and test a new hypothesis; to achieve different analytical objectives; to answer a fresh or propagated question; to improve the process execution results; or to resolve major problems in the process execution. **A feedback ring** is a small scale iteration that involves just two or more of the process phases. It involves the re-running of all phases inside the feedback ring. Feedback is usually triggered as a response to poor phase outcomes; problems with the running of the phase internal tasks; or due to the inadequacy of the phase prerequisites or inputs. **Rollback** is the undoing of the process execution. It requires preserving the current and the previous state of the process execution including the state of its phases running and the state of its deliveries generation. The **process termination** is usually decided in **process evaluation**, yet it is heavily influenced by **knowledge evaluation**.

Figure 14 provides seven possible process execution scenarios using a tree-like graph that demonstrates feedback, rollback, phase and process iteration mechanisms. **Scenario 1** illustrates normal process flow. **Scenario 2** illustrates feedback between the executed process phases due to poor model evaluation. **Scenario 3** illustrates rollback to undo a process execution feedback and resume the normal process execution after failing to build a better model. **Scenario 4** and **Scenario 5** illustrate examples for feedback conducted

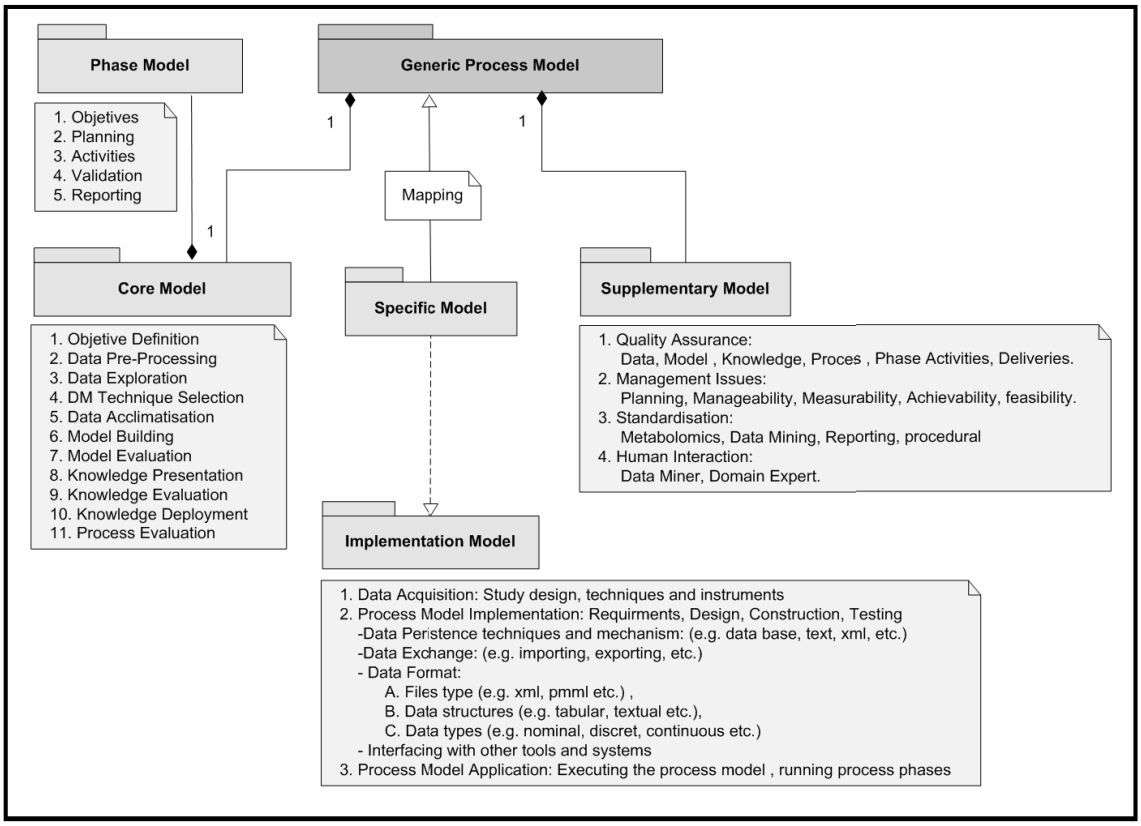

FIGURE 13. **Process Mapping-a UML Model showing the process mapping from generic to specific to implementation model.**

to produce alternative or better knowledge presentation and deployment respectively. **Scenario 6** illustrates an example of a process iteration due to setting new process objectives. Another possible reason for process iteration is changing the aims of the study. **Scenario 7** illustrates an example of a fresh process execution instantiation due to a new or significantly changed data.

H. DISCUSSION AND CONCLUSION

A novel knowledge discovery and data mining process model for metabolomics is introduced which has been implemented and applied to a number of datasets demonstrating its applicability to both data-driven and hypothesis-driven data mining approaches and its capability to achieve discovery-oriented, verification-oriented, predictive, or descriptive goals. Under it, a range of data mining tasks including regression, classification, rules induction segmentation, association, dimensionality reduction, correlation, hypothesis testing and feature extraction and analysis can be performed. The process provides a mechanism for defining feasible, achievable and measurable objectives that are based on matching the aims of metabolomics studies to data mining approaches, goals and tasks. It encourages the use of a strategy for selecting data mining techniques in a justifiable, and traceable way under good process management.

The process model satisfies the requirements of metabolomics data handling, pre-processing, pretreatment and interpretation. It addresses the shortcomings of the existing data process models and satisfies the requirements of metabolomics data mining. The process model is in line with the scientific nature of metabolomics investigations and is consistent with the cycle of knowledge [1]. It supports both deductive and inductive knowledge acquisition and also supports traceability, justifiability and reproducibility of data analysis procedures and results which are essential for scientific applications of data mining. Quality assurance is embedded in the design of the process model and all its phases. The model ensures the quality of its input and deliveries, the validity of its execution and the validity of the execution of the internal tasks in each of its phases. It evaluates built data mining models in the context of machine learning as well as in the context of metabolomics. Standardisation is also considered in the process model design and its execution. The process model terminologies are consistent with RSBI (ISA-TAB) [33], while the process flow is consistent with the MSI proposal [51]. In addition, an XML format is used for reporting the process execution and phase running, while PMML [130] is encouraged for reporting the data mining models.

The process model introduces major improvements and enhancements regarding process layout, emphasising the coherence of the process phases and including iteration, feedback and validation of the process flow. It introduces the concept of multiple evaluations in a data mining process

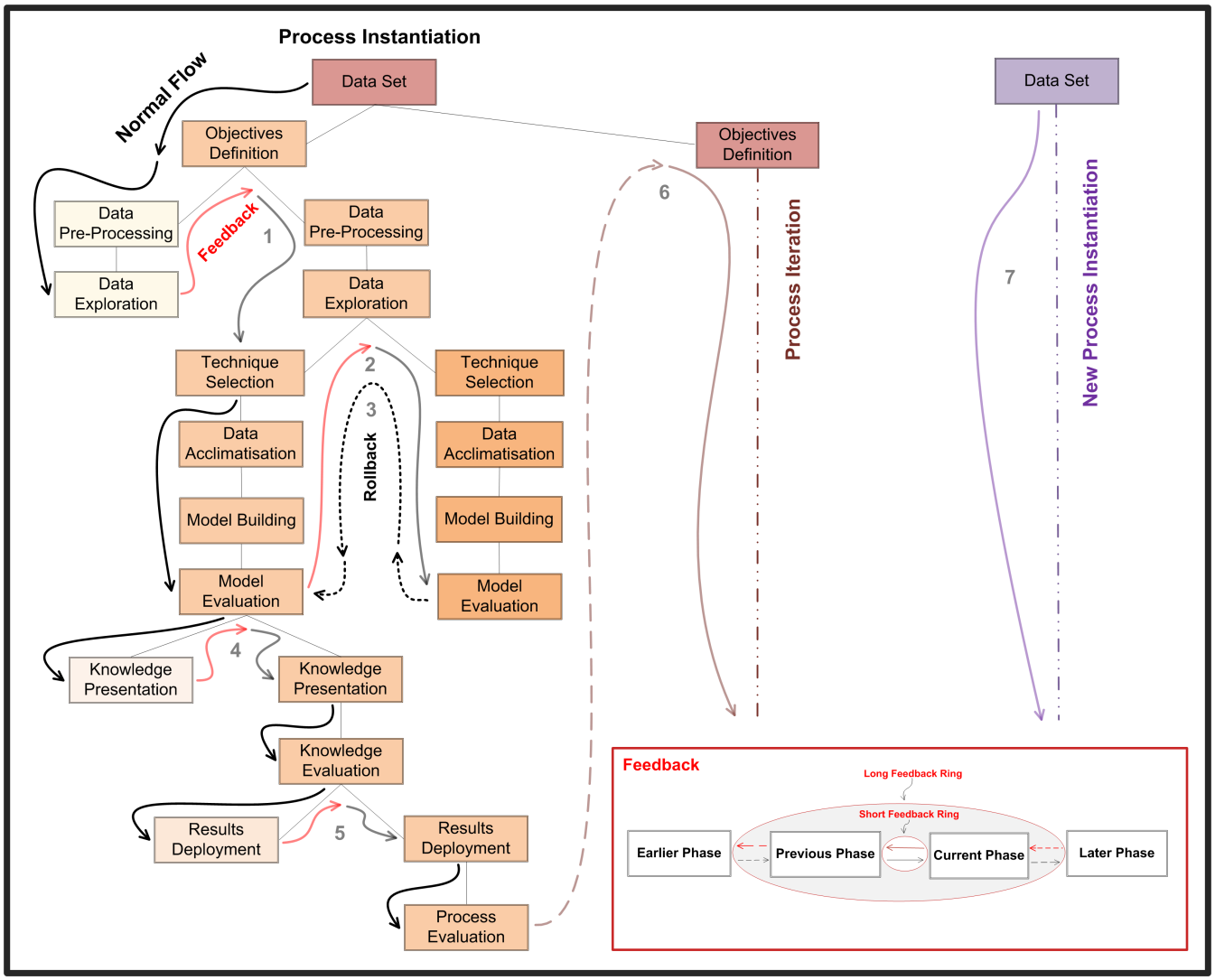

FIGURE 14. **Process Execution Scenarios- Solid black arrows illustrate normal process flow. Red arrows illustrate feedback between phases. Dotted black arrows represent rollbacks. Solid grey arrows illustrate resumed normal flow after a feedback. Dashed line illustrates process iteration, while purple arrows illustrate process execution instantiation. *The coloured rectangles* illustrates different versions produced during a phase running.**

where the data mining model is evaluated in its performance from a technical perspective and then the knowledge is evaluated in its contextual meaning as presented and interpreted metabolomics knowledge. Yet, the process also supports cross-delivery validation which was adapted from the software engineering V-model where the earlier deliveries of the process are checked for validity and consistency with corresponding deliveries of later phases in the process. Other desirable features include visualisation, data exploration, knowledge presentation and automation. It provides support for the practical aspects of metabolomics data mining which are embedded in the design of the process model layout and all of its phases. Human interaction is considered across all stages, both on the macro level of the process and its structure and on the micro level of its phases and their internal tasks. The process model structure and description provide details for planning and recording who is doing what, where in the process and how it is being done. Principles and best practices of project management are incorporated in every aspect of the proposed process. Management issues are considered on the level of the process and on the level of the tasks within phases such as activities planning, feasibility assessment, success definition and measurability in addition to resources management and allocation.

The proposed process model offers several contributions to metabolomics data analysis and to the design and automation of data mining process models. It introduces the concept of computer aided data mining, which aims not only to automate the various aspects of the data mining process (as others have suggested), but also to realise the layout, structure, and flow of the data mining process itself and to provide support for its various practical aspects. Furthermore, the features introduced in this model could be utilised to develop a generic scientific data mining process model that can be extended to

cover more scientific disciplines facilitated by the concept of mapping that was embedded in the process design and software implementation where the process model is broken down into groups of stages which consist of a set of the generic tasks that are customisable to a particular scientific application.

ACKNOWLEDGMENT

The authors thank Janet Taylor and Simon Garrett for their invaluable feedback and discussion during the research project.

REFERENCES

[1] D. B. Kell, "Metabolomics, machine learning and modelling: Towards an understanding of the language of cells," *Biochem. Soc. Trans.*, vol. 33, no. 3, pp. 520–524, Jun. 2005.

[2] A. BaniMustafa, "A knowledge discovery and data mining process model for metabolomics," Ph.D. dissertation, Dept. Comput. Sci., Aberystwyth Univ., Aberystwyth, Wales, 2012.

[3] A. B. Mendes, L. Cavique, and J. M. Santos, "Data mining process models: A roadmap for knowledge discovery," in *Quantitative Modelling in Marketing and Management*. Singapore: World Scientific, 2013, pp. 405–433.

[4] M. Rogalewicz and R. Sika, "Methodologies of knowledge discovery from data and data mining methods in mechanical engineering," *Manage. Prod. Eng. Rev.*, vol. 7, no. 4, pp. 97–108, Dec. 2016.

[5] U. Shafique and H. Qaiser, "A comparative study of data mining process models (KDD, CRISP-DM and SEMMA)," *Int. J. Innov. Sci. Res.*, vol. 12, no. 1, pp. 217–222, 2014.

[6] H. J. G. Palacios, R. A. J. Toledo, G. A. H. Pantoja, and Á. A. M. Navarro, "A comparative between CRISP-DM and SEMMA through the construction of a MODIS repository for studies of land use and cover change," *Adv. Sci., Technol. Eng. Syst. J.*, vol. 2, no. 3, pp. 598–604, Jun. 2017.

[7] A. Azevedo and M. F. Santos, "KDD, SEMMA and CRISP-DM: A parallel overview," in *Proc. ADIS Eur. Conf. Data Mining*, A. P. Abraham. Amsterdam, The Netherlands: IADIS, 2008, pp. 182–185.

[8] G. Mariscal, Ó. Marbán, and C. Fernández, "A survey of data mining and knowledge discovery process models and methodologies," *Knowl. Eng. Rev.*, vol. 25, no. 2, pp. 137–166, Jun. 2010.

[9] O. Niaksu, "Crisp data mining methodology extension for medical domain," *Baltic J. Modern Comput.*, vol. 3, no. 2, pp. 92–109, 2015.

[10] T. Li and D. Ruan, "An extended process model of knowledge discovery in database," *J. Enterprise Inf. Manage.*, vol. 20, no. 2, pp. 169–177, Feb. 2007.

[11] Ṁarbán, G. Mariscal, and J. Segovia, "A data mining & knowledge discovery process model," in *Data Mining and Knowledge Discovery in Real Life Applications*. Rijeka, Croatia: InTech, 2009.

[12] E Karišik, "A standardized data mining method in healthcare: A pediatric intensive care unit case study," M.S. thesis, Dept. Inf. Comput. Sci., Utrecht University, Utrecht, The Netherlands, 2018.

[13] Y. Li, M. A. Thomas, and K.-M. Osei-Bryson, "A snail shell process model for knowledge discovery via data analytics," *Decis. Support Syst.*, vol. 91, pp. 1–12, Nov. 2016.

[14] A. BaniMustafa and N. Hardy, "Applications of a novel knowledge discovery and data mining process model for metabolomics," 2019, *arXiv:1907.03755*. [Online]. Available: http://arxiv.org/abs/1907.03755

[15] A. BaniMustafa and N. Hardy, "Computer-aided data mining: Automating a novel knowledge discovery and data mining process model for metabolomics," 2019, *arXiv:1907.04318*. [Online]. Available: http://arxiv.org/abs/1907.04318

[16] A. BaniMustafa, *MeKDDaM-SAGA: A Software for Automating and Guiding a Knowledge Discovery and Data Mining Process Model for Metabolomics, Version 1.0*. Geneva, Switzerland: Zenodo, Jun. 2019, doi: 10.5281/zenodo.4015572.

[17] V. Maloney, "Plant metabolomics," *BioTeach J.*, vol. 2, no. 1, pp. 92–99, Fall 2004.

[18] K. Dettmer and D. Hammock, "Metabolomics: A new exciting field within the 'omics' sciences," *Evironmental Health Perspect.*, vol. 112, no. 7, pp. A396–A397, 2004.

[19] W. B. Dunn and D. I. Ellis, "Metabolomics: Current analytical platforms and methodologies," *Trends Anal. Chem.*, vol. 24, no. 4, pp. 285–294, 2005.

[20] K. Dettmer, P. A. Aronov, and B. D. Hammock, "Mass spectrometry-based metabolomics," *Mass Spectrometry Rev.*, vol. 26, no. 1, pp. 51–78, Jan. 2007.

[21] D. Kell, "Metabolomics and systems biology: Making sense of the soup," *Current Opinion Microbiology*, vol. 7, no. 3, pp. 296–307, Jun. 2004.

[22] R. Goodacre, S. Vaidyanathan, W. B. Dunn, G. G. Harrigan, and D. B. Kell, "Metabolomics by numbers: Acquiring and understanding global metabolite data," *Trends Biotechnol.*, vol. 22, no. 5, pp. 245–252, May 2004.

[23] V. Shulaev, "Metabolomics technology and bioinformatics," *Briefings Bioinf.*, vol. 7, no. 2, pp. 128–139, Mar. 2006.

[24] O. Fiehn, "Metabolomics—The link between genotypes and phenotypes," *Plant Mol. Biol.*, vol. 48, no. 1, pp. 155–171, 2002.

[25] J. Kopka, A. Fernie, W. Weckwerth, Y. Gibon, and M. Stitt, "Metabolite profiling in plant biology: Platforms and destinations," *Genome Biol.*, vol. 5, no. 6, p. 109.1–109.9, 2004.

[26] H. Fuell, "Options for the storage of the results of gas chromotography-mass spectrometry (GC-MS) analysis," Univ. Wales, Aberystwyth, U.K., Project Rep. g02006, 2003.

[27] W. Weckwerth and O. Fiehn, "Can we discover novel pathways using metabolomic analysis?" *Current Opinion Biotechnol.*, vol. 13, no. 2, pp. 156–160, Apr. 2002.

[28] J. Boccard, J.-L. Veuthey, and S. Rudaz, "Knowledge discovery in metabolomics: An overview of MS data handling," *J. Separat. Sci.*, vol. 33, no. 3, pp. 290–304, Feb. 2010.

[29] L. W. Sumner, P. Mendes, and R. A. Dixon, "Plant metabolomics: large-scale phytochemistry in the functional genomics era," *Phytochemistry*, vol. 62, no. 6, pp. 817–836, Mar. 2003.

[30] J. Xia, N. Psychogios, N. Young, and D. S. Wishart, "MetaboAnalyst: A Web server for metabolomic data analysis and interpretation," *Nucleic Acids Res.*, vol. 37, pp. W652–W660, Jul. 2009.

[31] H. Jenkins, H. Johnson, B. Kular, T. Wang, and N. Hardy, "Toward supportive data collection tools for plant metabolomics," *Plant Physiol.*, vol. 138, no. 1, pp. 67–77, May 2005.

[32] A. R. Jones *et al.*, "The functional genomics experiment model (FuGE): An extensible framework for standards in functional genomics," *Nature Biotechnol.*, vol. 25, no. 10, pp. 1127–1133, Oct. 2007.

[33] S.-A. Sansone *et al.*, "The first RSBI (ISA-TAB) workshop: 'Can a simple format work for complex studies,'" *OMICS, A J. Integrative Biol.*, vol. 12, no. 2, pp. 143–149, Jun. 2008.

[34] T. Ghosh, W. Zhang, D. Ghosh, and K. Kechris, "Predictive modeling for metabolomics data," in *Computational Methods and Data Analysis for Metabolomics*. New York, NY, USA: Springer, 2020, pp. 313–336.

[35] A. BaniMustafa, "Enhancing learning from imbalanced classes via data preprocessing: A data-driven application in metabolomics data mining," *ISC Int. J. Inf. Secur.*, vol. 11, pp. 79–89, Jan. 2019.

[36] S. Cardoso, T. Afonso, M. Maraschin, and M. Rocha, "WebSpecmine: A Website for metabolomics data analysis and mining," *Metabolites*, vol. 9, no. 10, p. 237, Oct. 2019.

[37] I. Martínez-Arranz, R. Mayo, M. Pérez-Cormenzana, I. Mincholé, L. Salazar, C. Alonso, and J. M. Mato, "Enhancing metabolomics research through data mining," *J. Proteomics*, vol. 127, pp. 275–288, Sep. 2015.

[38] A. Singh, "Tools for metabolomics," *Nature Methods*, vol. 17, no. 1, p. 24, 2020.

[39] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," *Commun. ACM*, vol. 39, no. 11, pp. 27–34, Nov. 1996.

[40] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "Knowledge discovery and data mining: Toward a unifying framework," in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining (KDD)*, E. Simoudis, J. Han, and U. Fayyad, Eds. Portland, OR, USA: AAAI Press, 1996, pp. 82–88.

[41] C. F. Taylor *et al.*, "Promoting coherent minimum reporting guidelines for biological and biomedical investigations: The MIBBI project," *Nature Biotechnol.*, vol. 26, no. 8, pp. 889–896, Aug. 2008.

[42] O. Maimon and L. Rokach, *Data Mining Knowl. Discovery Handbook*. New York, NY, USA: Springer, 2005.

[43] S. Sumathi and S. N. Sivanandam, "Data mining tasks, techniques, and applications," in *Introduction to Data Mining and its Applications* (Studies in Computational Intelligence). Berlin, Germany: Springer, 2006, pp. 195–216.

[44] R. Goodacre, *Data Analsysis Standards in Metabolomics*. Manchester, U.K.: Reserach Group Report, Univ. Manchester, 2006.
[45] A. H. BaniMustafa and N. W. Hardy, "A strategy for selecting data mining techniques in metabolomics," in *Plant Metabolomics: Methods Protocols* (Methods in Molecular Biology), vol. 860, N. W. Hardy and R. D. Hall, Eds. Clifton, NJ, USA: Springer, ch. 18, Feb. 2012, pp. 317–333.
[46] R. Goodacre, D. Broadhurst, A. Smilde, B. Kristal, J. Baker, R. Beger, C. Bessant, S. Connor, G. Capuani, A. Craig, T. Ebbels, D. Kell, C. Manetti, J. Newton, G. Paternostro, R. Somorjai, M. Sjostrom, J. Trygg, and F. Wulfert, "Proposed minimum reporting standards for data analysis in metabolomics," *Metabolomics*, vol. 3, no. 3, pp. 231–241, 2007.
[47] Thermo Scientific, "Introduction to Fourier transform infrared spectroscopy," Thermo Fisher Sci. Inc., Madison, WI, USA, Tech. Rep. BR50555_E 10/07M, 2007.
[48] I. Spasic, W. Dunn, G. Velarde, A. Tseng, H. Jenkins, N. Hardy, S. Oliver, and D. Kell, "MeMo: A hybrid SQL/XML approach to metabolomic data management for functional genomics," *BMC Bioinformatics*, vol. 7, no. 1, p. 281, 2006.
[49] L. W. Sumner *et al.*, "Proposed minimum reporting standards for chemical analysis," *Metabolomics*, vol. 3, no. 3, pp. 211–221, Sep. 2007.
[50] N. Hardy and H. Jenkins, "Reporting standards," in *Topics in Current Genetics*, vol. 18. Berlin, Germany: Springer, Jul. 2007, pp. 53–73.
[51] O. Fiehn, D. Robertson, J. Griffin, M. van der Werf, B. Nikolau, N. Morrison, L. W. Sumner, R. Goodacre, N. W. Hardy, C. Taylor, J. Fostel, B. Kristal, R. Kaddurah-Daouk, P. Mendes, B. van Ommen, J. C. Lindon, and S.-A. Sansone, "The metabolomics standards initiative (MSI)," *Metabolomics*, vol. 3, no. 3, pp. 175–178, Sep. 2007.
[52] R. D. Hall, "Plant metabolomics: From holistic hope, to hype, to hot topic," *New Phytologist*, vol. 169, no. 3, pp. 453–468, Feb. 2006.
[53] R. A. van den Berg, H. C. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf, "Centering, scaling, and transformations: Improving the biological information content of metabolomics data," *BMC Genomics*, vol. 7, no. 1, p. 142, Dec. 2006.
[54] S. Burke, "Missing values, outliers, robust statistics & non-parametric methods," LCGC Eur. Online Supplement, Cranbury, NJ, USA, Tech. Rep. 4, 2001.
[55] L. J. Sweetlove, R. L. Last, and A. R. Fernie, "Predictive metabolic engineering: A goal for systems biology," *Plant Physiol.*, vol. 132, no. 2, pp. 420–425, Jun. 2003.
[56] B. J. Read, "Data mining and science—Knwoledge discovery in science opposed to business," in *Proc. 12th ERCIM Workshop Database Res.* Didcot, U.K.: CLRC Rutherford Appleton Laboratory, Nov. 1999. [Online]. Available: https://www.ercim.eu/publication/ws-proceedings/12th-EDRG/
[57] C. Larman, *Agile Iterative Development: A Manager's Guide*. Reading, MA, USA: Addison-Wesley, 2004.
[58] M. Brown, W. B. Dunn, D. I. Ellis, R. Goodacre, J. Handl, J. D. Knowles, S. O'Hagan, I. Spasić, and D. B. Kell, "A metabolome pipeline: From concept to data to knowledge," *Metabolomics*, vol. 1, no. 1, pp. 39–51, Mar. 2005.
[59] S.-A. Sansone, D. Schober, H. J. Atherton, O. Fiehn, H. Jenkins, P. Rocca-Serra, D. V. Rubtsov, I. Spasic, L. Soldatova, C. Taylor, A. Tseng, and M. R. Viant, "Metabolomics standards initiative: Ontology working group work in progress," *Metabolomics*, vol. 3, no. 3, pp. 249–256, Sep. 2007.
[60] J. Ronald Brachman and T. Anand, "The process of knowledge discovery in databases: A first sketch," AAAI, Menlo Park, CA, USA, Tech. Rep. WS-94-03, 1994.
[61] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "CRISP-DM 1.0: Step-by-step data mining guide," SPSS Inc., Chicago, IL, USA, Tech. Rep. CRISPMWP-1104, 2000.
[62] C. Shearer, "The CRISP-DM model: The new blueprint for data mining," *J. Data Warehousing*, vol. 5, no. 4, pp. 13–22, 2000.
[63] K. J. W. R. W. Cios Pedrycz Swiniarski and L. A. Kurgan, *Data Mining: A Knowledge Discovery Approach*, 1st ed. New York, NY, USA: Springer, 2007.
[64] B. Dunkel, N. Soparkar, J. Szaro, and R. Uthurusamy, "Systems for KDD: From concepts to practice," *Future Gener. Comput. Syst.*, vol. 13, nos. 2–3, pp. 231–242, Nov. 1997.
[65] L. A. Kurgan and P. Musilek, "A survey of knowledge discovery and data mining process models," *Knowl. Eng. Rev.*, vol. 21, no. 1, pp. 1–24, Mar. 2006.
[66] R. Wirth and J. Hipp, "CRISP-DM—Towards a standard process model for data mining," in *Proc. 4th Int. Conf. Practical Appl. Knowl. Discovery Data Mining*, 2000, pp. 29–39.
[67] M. Ankerst, "Human involovment and interactivity of the next generation's data mining tools," in *Proc. Workshop Res. Issues Data Mining Knowl. Discovery Workshop Res. Issues Data Mining Knowl. Discovery*, Seattle, 2001, pp. 1–4.
[68] J. Han, L. V. S. Lakshmanan, and R. T. Ng, "Constraint-based, multidimensional data mining," *Computer*, vol. 32, no. 8, pp. 46–50, 1999.
[69] C. J. Hereth, S. Gerd, W. Rudolf, and W. Uta, "Conceptual knowledge discovery: A human-centred approache," *Appl. Artif. Intell. Int. J.*, vol. 17, no. 3, pp. 281–302, 2003.
[70] R. Brachman and T. Anand, "The process of knowledge discovery in data bases: A human centred approach," in *Proc. AKDDM*, Cambridge, MA, USA: MIT Press, 1996, pp. 37–58.
[71] C. Broeckling, T. Duran, and D. Huhman, "Metabolomics," Samuel Roberts Noble Found. Inc., Ardmore, OK, USA, Tech. Rep., May 2005.
[72] R. Pechter, "Conformance standard for the predictive model markup language," in *Proc. 4th Int. Workshop Data Mining Standards, Services Platforms (DMSSP)*. New York, NY, USA: ACM, 2006, pp. 6–13, doi: 10.1145/1289612.1289613.
[73] R. L. Grossman, "KDD workshop on data mining standards, services & platforms (DM-SSP) 2006," *ACM SIGKDD Explor. Newslett.*, vol. 8, no. 2, pp. 82–83, Dec. 2006.
[74] H. Jenkins *et al.*, "A proposed framework for the description of plant metabolomics experiments and their results," *Nature Biotechnol.*, vol. 22, no. 12, pp. 1601–1606, Dec. 2004.
[75] O. Fiehn, "Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks," *Comparative Funct. Genomics*, vol. 2, no. 3, pp. 155–168, 2001.
[76] T. Bamba and E. Fukusaki, "Technical problems and practical operations in plant metabolomics," *J. Pesticide Sci.*, vol. 31, no. 3, pp. 300–304, 2006.
[77] W. Trochim and J. onnelly, *The Research Methods Knowledge Base*, 3rd ed. Cincinnati, OH, USA: Atomic Dog Publishing, 2007.
[78] M. Goebel and L. Gruenwald, "A survey of data mining and knowledge discovery software tools," *ACM SIGKDD Explor. Newslett.*, vol. 1, no. 1, pp. 20–33, Jun. 1999.
[79] S. Ryszad Michalski, I. Bratko, and M. Kubat, *Machine Learning and Data Mining- Methods and Applications*. Hoboken, NJ, USA: Wiley, 1998.
[80] K. Pelckmans, J. De Brabanter, J. A. K. Suykens, and B. De Moor, "Handling missing values in support vector machine classifiers," *Neural Netw.*, vol. 18, nos. 5–6, pp. 684–692, Jul. 2005.
[81] J. Xi, "Outlier detection algorithms in data mining," in *Proc. 2nd Int. Symp. Intell. Inf. Technol. Appl.*, Dec. 2008, pp. 94–97.
[82] R. L. Grossman. (Sep. 2004). *Event Based Data Mining Process Models*. Open Data Partners, pp. 1–3. Accessed: Jul. 5, 2010. [Online]. Available: http://www.opendataresearch.com/ebpm.htm
[83] *Introduction to Data Mining and Knowledge Discovery*, Technology Report, Two-Crows Corporation, Potomac, MD, USA, 1999.
[84] D. A. Keim, "Information visualization and visual data mining," *IEEE Trans. Vis. Comput. Graphics*, vol. 8, no. 1, pp. 1–8, Aug. 2002.
[85] M. Ankerst, "Visual data mining with pixel-oriented visualization techniques," in *Proc. ACM SIGKDD Workshop Vis. Data Mining*, 2001, p. 23.
[86] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms*. Hoboken, NJ, USA: Wiley, 2003.
[87] M. Pedro, "Emerging bioinformatics for the metabolome," *Briefings Bioinf.*, vol. 3, no. 2, pp. 134–145, Jan. 2002.
[88] M. M. Campos, P. J. Stengard, and B. L. Milenova, "Data-centric automated data mining," in *Proc. 4th Int. Conf. Mach. Learn. Appl. (ICMLA)*, 2005, pp. 97–104.
[89] M. R. Syed and S. N. Syed, *Handbook of Research on Modern Systems Analysis and Design Technologies and Applications*. Hershey, PA, USA: Information Science Reference, 2009.
[90] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," *Informatica*, vol. 31, no. 1, pp. 249–268, 2007.
[91] J. G. Williams and Z. Huang, "Modeling the KDD process—A four stage and four elements model," Division Inf. Technol., CSIRO, Canberra, ACT, Australia, Portfolio Rep. TR DM 96013, Feb. 1996 1996.
[92] N. Jovanovic, V. Milutinovic, and Z. Obradovic, "Foundations of predictive data mining," in *Proc. 6th Seminar Neural Netw. Appl. Electr. Eng.*, Sep. 2002, pp. 53–58.

[93] Nautilus Systems, Inc. (2005). *The Data Mining Process*. [Online]. Available: http://www.nautilus-systems.com/process.html
[94] H. A. Milley, D. J. Seabolt, and S. J. Williams, "Data mining and the case for sampling solving business problems using SAS enterprise miner software," SAS Inst. Inc., Cary, NC, USA, Tech. Rep. 19963US.0399 REV, 1998. [Online]. Available: https://sceweb.uhcl.edu/boetticher/ML_DataMining/SAS-SEMMA.pdf
[95] D. J. Hand, Heikki Mannila, and P. Smyth, *Principles of Data Mining Adaptive Computation and Machine Learning*. Cambridge, MA, USA: MIT Press, 2001.
[96] C. Rolland, "A comprehensive view of process engineering," in *Proc. 10th Int. Conf. Adv. Inf. Syst. Eng. (CAiSE)* Berlin, Germany: Springer-Verlag, 1998, pp. 1–24.
[97] B. Warboys, D. Avrillionis, R. Conradi, P.-Y. Cunin, M. N. Nguyen, and I. Robertson, "Meta-process," in *Softw. Process: Princ., Methodology, Technol.*, pp. 53–94. Springer-Verlag, Berlin, Heidelberg, 1999.
[98] W. Scacchi, "Process models in software engineering," in *Encyclopedia of Software Engineering*, J.J. Marciniak, Ed. 2n ed. New York, NY, USA: Wiley, 2001.
[99] I. Sommerville, "Software engineering," in *International Computer Science Series*, 8th ed. Reading, MA, USA: Addison-Wesley, 2007.
[100] A. Fuggetta, "Software process: A roadmap," in *Proc. Conf. The Future Softw. Eng.*, Limerick, Ireland, 2000, pp. 25–34.
[101] V. Schuppan and W. Rußwurm, "A CMM-based evaluation of the V-model 97," in *Software Process Technology*, vol. 1780, R. Conradi, Ed. Berlin, Germany: Springer, 2000, pp. 69–83.
[102] R. Constantinescu, "V-model role engineering," *Inf. Economica*, vol. 13, no. 1, pp. 38–46, 2009.
[103] A. Roy Boggs, "The SDLC and six sigma: An essay on which is which and why," *Issues Inf. Syst.*, vol. 5, no. 1, pp. 36–42, 2004.
[104] P. Langley and H. A. Simon, "Applications of machine learning and rule induction," *Commun. ACM*, vol. 38, no. 11, pp. 54–64, Nov. 1995.
[105] G. Dodig-Crnkovic, "Scientific methods in computer science," in *Proc. Conf. Promotion Res. IT New Universities at Univ. Colleges in Sweden*, Skövde, Sweden: Mälardalen Univ., Dept. Comput. Sci., 2002, pp. 126–130.
[106] S. Carrol and D. Goodstein, "Defining the scientific method," *Nature Methods*, vol. 6, no. 4, p. 237, 2009.
[107] R. Gorini, "Al-haytham the man of experience, first steps in the science of vision," *J. Int. Soc. Hist. Islamic Med.*, vol. 2, no. 4, pp. 53–56, 2003.
[108] P. Panov, S. Dzeroski, and L. Soldatova, "OntoDM: An ontology of data mining," in *Proc. Data Mining Workshops, Int. Conf.*, Dec. 2008, pp. 752–760.
[109] T. R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing?" *Int. J. Hum.-Comput. Stud.*, vol. 43, nos. 5–6, pp. 907–928, Nov. 1995.
[110] S. D. P. Panov and N. L. Soldatova, "Towards an ontology of data mining investigations," in *Proc. 12th Intl. Conf. Discovery Sci.*, 2009, pp. 257–271.
[111] C. Diamantini, D. Potena, and E. Storti, "Ontology-driven KDD process composition," in *Proc. 8th Int. Symp. Intell. Data Anal.*, N. Adams, Ed. Berlin, Germany: Springer, 2009, pp. 285–296.
[112] A. Bernstein, F. Provost, and S. Hill, "Toward intelligent assistance for a data mining process: An ontology-based approach for cost-sensitive classification," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 503–518, Apr. 2005.
[113] A. Kalousis, A. Bernstein, and M. Hilario, "Meta-learning with kernels and similarity functions for planning of data mining workflows," in *Proc. Planing Learn Workshop (PlanLearn)*, Helsinki, Finnland, 2008, pp. 23–28.
[114] M. Žáková, P. Křemen, F. Železnỳ, and N. Lavrač, "Planning to learn with a knowledge discovery ontology," in *Planning to Learn Workshop (PlanLearn 2008) at ICML*, P. Brazdil, A. Bernstein, and L. Hunter, Eds. Helinki, Finland, Jul. 2008.
[115] N. W. Hardy and C. F. Taylor, "A roadmap for the establishment of standard data exchange structures for metabolomics," *Metabolomics*, vol. 3, no. 3, pp. 243–248, Sep. 2007.
[116] H. Neuweger, S. P. Albaum, M. Dondrup, M. Persicke, T. Watt, K. Niehaus, J. Stoye, and A. Goesmann, "MeltDB: A software platform for the analysis and integration of metabolomics experiment data," *Bioinformatics*, vol. 24, no. 23, pp. 2726–2732, Dec. 2008.
[117] L. N. Soldatova and R. D. King, "An ontology of scientific experiments," *J. Roy. Soc. Interface*, vol. 3, no. 11, pp. 795–803, Dec. 2006.
[118] I. Niles and A. Pease, "Towards a standard upper ontology," in *Proc. Int. Conf. Formal Ontology Inf. Syst. FOIS*, 2001, pp. 2–9.
[119] U. Johansson, L. Niklasson, and R. Knig, "Accuracy vs. Comprehensibility in data mining models," in *Proc. 7th Int. Conf. Inf. Fusion*, Stockholm, Sweden, 2004, pp. 295–300.
[120] H. Kerzner, *Project Management: A Systems Approach to Planning, Scheduling, and Controlling*, 8th ed. New York, NY, USA: Wiley, 2003.
[121] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophys.*, vol. 5, no. 4, pp. 115–133, Dec. 1943.
[122] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biol. Cybern.*, vol. 43, no. 1, pp. 59–69, 1982.
[123] V. Vapnik, "Principles of risk minimization for learning theory," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 4, J. E. Moody, S. J. Hanson, and R. Lippmann, Eds. San Mateo, CA, USA: Morgan Kaufmann, 1992, pp. 831–838.
[124] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th Annu. Workshop Comput. Learn. Theory - COLT*, 1992, pp. 144–152.
[125] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *London, Edinburgh, Dublin Phil. Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, Nov. 1901.
[126] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, Sep. 1967.
[127] M. Barker and W. Rayens, "Partial least squares for discrimination," *J. Chemometrics*, vol. 17, no. 3, pp. 166–173, 2003.
[128] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.
[129] T. Kam Ho, "Random decision forests," in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, Aug. 1995, pp. 278–282.
[130] A. Guazzelli, M. Zeller, W.-C. Lin, and G. Williams, "PMML: An open standard for sharing models," *R J.*, vol. 1, no. 1, pp. 60–65, 2009.

AHMED BANIMUSTAFA (Member, IEEE) is currently the Head of the Department of Software Engineering, ISRA University. His major research interests include data mining, data science, machine learning, software engineering, bioinformatics, and metabolomics. He received the Ph.D. degree in computer science from the University of Wales, Aberystwyth (Aberystwyth University), U.K., in 2012, where he was a member of the Computational Biology Research Group. He joined ISRA University in 2019 and worked earlier as an Assistant Professor with the American University of Madaba from 2013 to 2019; a Part-Time Tutor and a Lecturer at Aberystwyth University from 2007 to 2011; and a Lecturer at Philadelphia University from 2003 to 2006. He also worked as a Part-Time Lecturer at Jordan University of Science and Technology (JUST) in 2003, 2004, and 2013, respectively, and spent a six-month internship at Motorola Mobility, Swindon, U.K., in 2002 (Acquired later by Google LLC). He is a member of the ACM and JSSR and was also a member of Metabolomics Society.

NIGEL HARDY is currently a Lecturer with the Department of Computer Science, Aberystwyth University, with an interest in bioinformatics, software engineering, and database systems. He has been a member of the Bioinformatics and Computational Biology Research Group and a Senate Representative of the Department of Computer Science. His research interest concentrates on the applications of data handling techniques in biology, specifically in metabolomics. His former research interest is focused on robot error detection and the abstraction of sensor usage to support rapid reconfiguration and reprogramming of systems with significant numbers of transducers. He was a founding member of the Metabolomics Society and also a member of the oversight committee of the MSI and the Co-Chair of the Exchange Format Workgroup.